

Ultrahigh-Throughput Microfluidic Droplet Screening of Metagenomic Libraries for Esterases and Kemp Eliminases



Philip Mair

This dissertation is submitted for the degree of
Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Philip Mair
October 2019

Ultrahigh-Throughput Microfluidic Droplet Screening of Metagenomic Libraries for Esterases and Kemp Eliminases

Philip Mair

In the search for new enzymes, functional metagenomic libraries are particularly interesting since they give access to the genomes of the 99% of microorganisms which cannot be cultured in the laboratory. However, using classical biochemical assays to screen such libraries is impeded by the fact that enzymes for a given reaction are incredibly rare (1 in 10^5 on average). Ultrahigh-throughput droplet microfluidics has emerged as an effective technology to overcome this limitation, with the ability to screen 10^7 reactions per day. However, fewer than ten enzyme assays have been implemented in droplets to date and only once has droplet microfluidics been applied to functional metagenomics. Therefore, for the wider adoption of this technology, it is critical to develop more enzyme assays in droplets.

To address this need, first I built an ultrahigh-throughput droplet sorting instrument and then set out to establish two new enzyme assays in droplets: one for the industrially-important esterase reaction and another for the Kemp elimination, an artificial reaction.

In Chapter 2, I describe the state-of-the-art fluorescence-activated droplet sorter (FADS) I developed for use with fluorogenic enzyme substrates. I improved this instrument incrementally based on the needs of collaborative projects, which ensured the success of these projects in screening enzyme libraries.

In Chapter 3, I established an esterase assay and performed the first reported functional metagenomic screen for esterases in droplets. Over 30 million droplets were sorted, amounting to $10\times$ coverage of a metagenomic library consisting of over one million members; making this the largest esterase screen performed to date. Twelve clones encoding thirteen novel esterases were isolated. The majority were members of the α/β -hydrolase super-family of proteins. Four came from small families taking up less than 1% of the sequence space within the super-family. Therefore, functional screening at ultrahigh-throughput provided access to thinly populated sequence-space. It is unlikely that these sequences would have been explored using prediction-based methods. Eight out of eleven enzymes had detectable thio-esterase activity, three had β -lactamase activity, one had β -galactosidase activity, and one had Kemp eliminase activity. This finding corroborates the idea that the ability of enzymes to catalyse more than one reaction, a property called enzyme promiscuity, is widespread.

In Chapter 4, I established a Kemp eliminase assay in droplets with the aim of exploring how widespread promiscuous enzymes catalysing this non-natural reaction are. Using the substrate 5-nitro-1,2-benzisoxazole in combination with absorbance-activated droplet sorting (AADS). I describe the enrichment of a previously reported Kemp eliminase, HG3.17,

over a negative control using droplet microfluidics and use this method to screen substitution, insertion and deletion libraries of HG3.17. Active library variants were enriched and the variants with improved soluble expression isolated. Five locations were identified that tolerate insertions and deletions and, in one investigated case, have improved soluble expression. These variants may serve as starting points to explore previously inaccessible mutational trajectories to improve the catalytic parameters of HG3.17. Due to the limited sensitivity of the assay, functional metagenomic screening was not possible using this substrate.

In Chapter 5, I report the newly-discovered fluorogenic Kemp substrate 5-azido-1,2-benzisoxazole, reported and characterised here for the first time. I established this substrate in droplets and, using the FADS instrument, enriched the Kemp eliminase HG3.17 over a negative control. This assay was sufficiently sensitive to detect Kemp eliminase activity in a metagenomic library. The apparent hit rate was comparable to that of the esterases, suggesting that promiscuous enzymes capable of catalysing this reaction are commonplace. A large number of false positives impeded the straightforward isolation of the responsible library members. This constraint is likely to be overcome in future by using a bias-free metagenomic library. Here, a combination of functional re-screening and sequencing allowed the identification of one lead: a predicted class IV adenylyl cyclase.

In conclusion, I have built a highly sensitive droplet sorting instrument using which I isolated thirteen new esterases, established the first Kemp eliminase assay in droplets, used both for mutant library and metagenomic screening, and additionally enabled the improvement of numerous enzymes through library screenings in collaboration with others. This work contributes to the success of ultrahigh-throughput droplet microfluidics in furthering our understanding of enzymes in nature and our ability to tailor them for green chemistry applications in industry.

To my parents

1. Die Welt ist alles, was der Fall ist.
2. Was der Fall ist, die Tatsache, ist das Bestehen von Sachverhalten.
3. Das logische Bild der Tatsachen ist der Gedanke.
4. Der Gedanke ist der sinnvolle Satz.
5. Der Satz ist eine Wahrheitsfunktion der Elementarsätze.
(Der Elementarsatz ist eine Wahrheitsfunktion seiner selbst.)
6. Die allgemeine Form der Wahrheitsfunktion ist: $[p, \xi, N(\xi)]$.
Dies ist die allgemeine Form des Satzes.
7. Wovon man nicht sprechen kann, darüber muss man schweigen.

Tractatus Logico-Philosophicus, Ludwig Wittgenstein

Part of the work presented here was published in:

1. Mair, P., Gielen, F. & Hollfelder, F. Exploring sequence space in search of functional enzymes using microfluidic droplets. *Curr. Opin. Chem. Biol.* 37, 137–144 (2017).
2. Gielen, F., Colin, P. Y., Mair, P. & Hollfelder, F. Ultrahigh-throughput screening of single-cell lysates for directed evolution and functional metagenomics. In *Methods in Molecular Biology* (eds. Bornscheuer, U. T. & Höhne, M.) 1685, 297–309 (Springer New York, 2018).

And submitted for publication and currently available at:

3. Loo, B. van, Heberlein, M. Mair, P., Zinchenko, A., Schuurmann, J., Eenink, B.D.G., Dilkate, C., Jose, J., Hollfelder, F. & Bornberg-Bauer, E. High-Throughput, Lysis-free Screening for sulfatase Activity Using *Escherichia coli* Autodisplay in Microdroplets. *bioRxiv* 479162 (2018). doi:10.1101/479162

Acknowledgements

I would like to thank my supervisor, Florian Hollfelder, for the opportunity to work on these exciting projects as well as the many ideas, discussions, encouragement, and funding to make it all happen. I am also grateful to the EPSRC Doctoral Training Program in Sensor Technologies and Applications (EP/L015889/1) to provide the funding for my master and doctoral studies.

How could I not start with Liisa Van Vliet, my first mentor in the Hollfelder group, who introduced me to microfluidics and always supported me and my work with her unbreakable optimism and can-do attitude. I would like to thank Fabrice Gielen whom I'm indebted to for my initial training in the group and who always offered his advice, guidance, and support on everything in electronics, microfluidics, academia, and beyond. I am deeply grateful to Mariana Rangel Pereira and Josephin Holstein for the close collaboration over the last few years and for offering their expertise, help, advice, company in the deep darkness of the laser cave, and most importantly their friendship.

It is hard to summarise all the support and encouragement I received from the members of the Hollfelder group. A special mention should go to Stéphane Emond for his advice in molecular biology, Christian Gylstorff for his ingenious mini-inventions such as the infamous Gylstorff chamber, Thomasz Kaminski for simply making things happen with stunning efficiency, and Stefanie Neun, Paul Zurek, David Schnettler Fernández for inspiring scientific discussions. I have to thank Raphaëlle Hours for sharing a glorious moment in demonstrating droplet sorting to Frances Arnold. I am also grateful to Pierre-Yves Colin for introducing me to the biochemical workflows in droplet screening and Anastasia Zinchenko for being living proof that anything can be achieved with hard work and determination. I am also thankful to Part II student Abi Turner, who supported us in measuring some of the esterase kinetics, and to Ana Torrado Agrasar for the gift of fluorescein-dihexanoate. I also thank Bert Van Loo and Magdalena Heberlein for the opportunity to work on the autodisplay project.

How could I have achieved any of this without the support of my friends near and far, my fellow rowers at Churchill College Boatclub, and my incredible housemates and friends Frabi and Belsa at our little Eachard Estate.

Ohne die Unterstützung meiner Eltern, Karin und Roman, wäre ich nie überhaupt bis zur Schwelle von Cambridge gekommen. Eure Liebe und euer Verständnis haben mich zu dem gemacht, der ich heute bin.

Enfin, j'aimerais te remercier toi, Chris, pour m'avoir soutenu, aidé, conseillé, écouté, consolé, amusé et simplement aimé.

Table of contents

List of figures	xxi
List of tables	xxv
Acronyms	xxvii
1 Introduction	1
1.1 Enzymes	1
1.1.1 How to find an enzyme	2
1.2 Metagenomics	5
1.2.1 Early studies and metagenomic sequencing	5
1.2.2 Functional metagenomic screening	6
1.3 Droplet microfluidics	8
1.3.1 Screening enzyme libraries using droplet microfluidics	11
1.3.2 Enzyme substrates in droplet experiments	14
1.3.3 Enzyme assays available in microfluidic droplets	15
1.4 Ultrahigh-throughput metagenomics in droplets	18
1.4.1 Goals of this thesis	20
2 The fluorescence-activated droplet sorter	23
2.1 Abstract	23
2.2 Introduction	24
2.3 Description of the FADS set-up	27
2.3.1 The optical and electronic set-up	27
2.3.2 The sorting algorithm	28
2.4 Results and discussion	36
2.4.1 Single-photon counting module <i>versus</i> photomultiplier tube	36
2.4.2 Improvements to the FADS were driven by research projects	40
2.5 Conclusions	42

3	Functional metagenomic screening for esterases in droplets	45
3.1	Abstract	45
3.2	Introduction	46
3.3	Establishing the esterase assay in droplets	53
3.4	Screening of the SCV Library	59
3.4.1	Droplet screening and hit recovery	59
3.4.2	Sequence analysis of the selected clones	63
3.5	Quantitative analysis of the newly identified esterases	71
3.5.1	Esterase kinetics with p-nitrophenyl carboxylates	71
3.5.2	Melting temperatures	74
3.5.3	Screening for promiscuous reactions	75
3.6	Conclusion	80
4	An absorbance-activated droplet sorting assay for the Kemp elimination	83
4.1	Abstract	83
4.2	Introduction	84
4.2.1	The Emergence of Phosphotriesterases	85
4.2.2	How To Find Starting Points of Evolution	85
4.2.3	The Kemp Elimination	86
4.3	Overview of this Chapter	94
4.4	The absorbance-based Kemp eliminase assay	98
4.4.1	Establishment of buffer conditions and enzyme controls	98
4.4.2	The Kemp reaction product exchanges between droplets	100
4.4.3	Modified 1,2-benzisoxazole substrates	105
4.4.4	In-line droplet generation and sorting	106
4.4.5	Activity of HG3.17 is detectable in droplets	110
4.4.6	Enrichment of HG3.17 using AADS is efficient	111
4.4.7	No activity observed in functional metagenomic screening	113
4.5	Directed evolution of HG3.17 using AADS	115
4.5.1	The screening workflow	118
4.5.2	Directed evolution of HG3.17 using substitutions or InDels	119
4.6	Conclusions	147
5	A novel fluorogenic Kemp Substrate	149
5.1	Abstract	149
5.2	A moment of serendipity: a fluorogenic Kemp substrate	150
5.2.1	Reproducing the initial observations	153

5.2.2	Two reaction steps lead to fluorescence	157
5.2.3	NMR and Mass spectroscopy of the reaction products	161
5.2.4	The single-cell lysate assay yields green fluorescence	166
5.2.5	The fluorophore emits at 335, 544, and 600 nm	168
5.2.6	Enrichment of HG3.17 using substrate 6a	172
5.3	Functional metagenomic screening for Kemp eliminases	173
5.3.1	Screening of the metagenomic SCV library	173
5.3.2	Sequence analysis and origin of the false positives	176
5.3.3	Droplet screening using the small SCV library reduced false positive rate	179
5.3.4	Re-Screening using absorbance as primary assay readout	180
5.3.5	Combining absorbance and sequencing data yields a potential hit .	181
5.4	Conclusions	188
6	Conclusions	191
7	Methods	199
7.1	Microfluidics	199
7.1.1	Fabrication of microfluidic devices	199
7.1.2	Generation and incubation of droplets	200
7.1.3	Droplet leakage assays	200
7.1.4	Fluorescence-activated droplet sorting (FADS)	201
7.1.5	Absorbance-activated droplet sorting (AADS)	202
7.1.6	Droplet conditions for the metagenomic esterase screen	202
7.1.7	Droplet conditions for the Kemp eliminase assays	202
7.1.8	DNA Recovery from microfluidic droplets	203
7.2	Biochemistry	204
7.2.1	General cloning procedures	204
7.2.2	Preparation of electro-competent cells	204
7.2.3	Construction of the metagenomic SCV library	205
7.2.4	Construction of the epPCR Library of HG3.17	205
7.2.5	Construction of the deletion library of HG3.17	206
7.2.6	Construction of the insertion library of HG3.17	207
7.2.7	Expression of HG3.17 libraries for droplet screening	208
7.2.8	Re-screening of HG3.17 variants in 96-well plates	208
7.2.9	Protein expression and purification	209
7.2.10	p-Nitrophenyl ester kinetics	210

7.2.11	Differential scanning fluorimetry	210
7.2.12	Catalytic promiscuity test	210
7.2.13	Kemp elimination kinetics	211
7.3	Chemistry	211
7.3.1	Synthesis of 5-nitro-1,2-benzisoxazole 2a	211
7.3.2	Synthesis of 4-nitro-2-cyanophenol 2b	211
7.3.3	Synthesis of 5-amino-1,2-benzisoxazole 5a	212
7.3.4	Synthesis of 5-azido-1,2-benzisoxazole 6a	212
7.3.5	Small molecule characterisation	212
7.4	Sequence Similarity Networks	213
Bibliography		215
Appendix A Supplementary Data Chapter 2		239
Appendix B Supplementary Data Chapter 3		243
B.1	DNA Inserts	243
B.1.1	N1 DNA Insert	243
B.1.2	N2 DNA Insert	246
B.1.3	N7 DNA Insert	247
B.1.4	N11 DNA Insert	249
B.1.5	N13 DNA Insert	249
B.1.6	N16 DNA Insert	251
B.1.7	N18 DNA Insert	252
B.1.8	N20 DNA Insert	253
B.1.9	N26 DNA Insert	253
B.1.10	N33 DNA Insert	254
B.1.11	RR11 DNA Insert	255
Appendix C Supplementary Data for Chapter 4		265
C.1	Amino-acid sequence of HG3.17	265
Appendix D Supplementary Data for Chapter 5		273

List of figures

1.1	Sources of new enzymes	3
1.2	Construction of a metagenomic library	7
1.3	Droplet Making	9
1.4	The Poisson Distribution	10
1.5	Photolithography and surface modification	12
1.6	General droplet workflow of library screening	13
1.7	Workflow to establish a droplet screening assay	21
2.1	Dielectrophoresis	25
2.2	FADS Set-up	29
2.3	Simulated droplet signal	31
2.4	Flowchart FPGA	32
2.5	Flowchart DMA Engine	34
2.6	Flowchart LabView	35
2.7	SPCM vs APD Detector	38
2.8	Linearity of PMT measurement	39
2.9	Measurement of signal amplitude and width	41
2.10	Width versus FWHM as a measure of droplet size	43
3.1	Applications of lipases and esterases	47
3.2	Mechanism of ester hydrolysis	48
3.3	Conventional phenotypic esterase screens	49
3.4	Workflow	52
3.5	Plate assay with esterase controls	54
3.6	Plate assay with esterase controls after 18 h	55
3.7	Droplets containing an esterase control	56
3.8	Histogram of droplets containing a positive esterase control	57
3.9	Histograms of different controls in droplets	58

3.10	Workflow of the functional metagenomic screen	59
3.11	Histograms of the droplet sorting	60
3.12	Colonies on tributyrin with halos	62
3.13	Confirmation of activity with fluorescein dihexanoate	62
3.14	Esterase ORFs in the DNA inserts	65
3.15	Sequence similarity network of α/β -hydrolase superfamily	68
3.16	The α/β -hydrolase fold	70
3.17	Example of kinetic parameter determination	72
3.18	Catalytic efficiency depending on ester chain length	73
3.19	Catalytic parameters of the metagenomic hits	73
3.20	Melting Temperatures	74
3.21	Catalytic promiscuity test	76
3.22	Promiscuous Activities of N1O5 and N7 and Cross-inhibition	79
4.1	The Kemp elimination	83
4.2	Factors contributing to the catalysis of the Kemp reaction	87
4.3	Catalysts of the Kemp elimination	88
4.4	Comparison of the second order rate constants of Kemp catalysts	90
4.5	Proposed mechanism for the Kemp elimination in heme-containing enzymes	93
4.6	Principle of absorbance-activated droplet sorting	95
4.7	Overview of the results	97
4.8	General base catalysis of the Kemp Elimination in seven buffers	99
4.9	Positive and negative controls for the Kemp Elimination in well-plate	101
4.10	Droplet Generation for Leakage Analysis	101
4.11	Analysis of reaction product leakage in droplets	103
4.12	Leakage Timecourses	104
4.13	Kemp Products tested for droplet leakage	106
4.14	Design of the Inline AADS	108
4.15	HG3.17 activity in droplets	110
4.16	Enrichment of Kemp eliminase HG3.17	112
4.17	Colony PCR of HG3.17 Enrichment	113
4.18	Histogram of the metagenomic library SCV using AADS	114
4.19	Possible sequence changes in the HG3.17 libraries	115
4.20	Construction of the HG3.17 libraries	117
4.21	Screening of HG3.17 Libraries	118
4.22	Mutations in naive epPCR library	120
4.23	Histograms of HG3.17 epPCR Library Sort	122

4.24	cell lysate assay in plates	124
4.25	Histogram of initial rates pre- and post-sorting	125
4.26	Histogram of initial rates pre- and post-sorting	126
4.27	Histograms of HG3.17 epPCR Library Sort	128
4.28	Initial rates in cell lysate before and after sorting	129
4.29	Location of mutations in the structure of HG3.17	130
4.30	Soluble Expression of HG3.17 variants	132
4.31	Mutations in naive InDel libraries	133
4.32	Historgrams of the InDel library droplet screening	133
4.33	Empirical cumulative probability functions of the three libraries	134
4.34	Initial rates in cell lysate before and after sorting InDel libraries	136
4.35	Soluble expression of deletion variant 4A09	137
4.36	Sequencing results deletion plate	139
4.37	Initial rates in cell lysate before and after sorting InDel libraries	140
4.38	InDel activity relative to amino-acid position in HG3.17	141
4.39	Location of InDels in HG3.17 structure	145
4.40	Detail of HG3.17 structure	146
5.1	Synthesis of substrates 6a and 7a	151
5.2	Discovery of Two Fluorogenic Kemp Substrates	152
5.3	Turnover of fluorogenic Kemp substrates by HG3.17	155
5.4	Exposure to UV light is necessary to form the red fluorophore	156
5.5	Two-step mechanism leads to fluorescence	157
5.6	Dark and UV reaction of 6a with HG3.17	158
5.7	Dependence of red fluorescence emission on pH	160
5.8	Possible reaction pathways of 6a	161
5.9	Proton NMR of the dark and UV exposed 6a reaction products	163
5.10	Elution profiles of 6a and the two reaction products	165
5.11	Mass spectrum of the final product	167
5.12	The product emits green light in droplets	168
5.13	The influence of cell lysate on emission	170
5.14	Red and green emission in different solvents	171
5.15	Histogram of HG3.17 enrichment with 6a	173
5.16	Histograms of the metagenomic sort using 6a	175
5.17	Result of the plate screen with 6a	177
5.18	Location of deletions in pZero2	178
5.19	Histogram of small SCV library droplet sort	180

5.20	384-well controls for re-screening of the small SCV library	182
5.21	Re-screening of small SCV library using absorbance assay	183
5.22	Different functional readouts of one plate	184
5.23	Insert Size Distribution	185
5.24	Absorbance assay data combined with sequencing information	185
5.25	Structural model of G06	187
6.1	Possible sources of error and bias	195
A.1	Spectra of filters in optical set-up	239
A.2	2D Data of linearity measurement	240
A.3	Simulation of fluorescence signal and measurements	241
B.1	pZero2 Vector Map	244
B.2	pHAT2 Vector Map	258
B.3	Sequence similarity network SGNH superfamily	259
B.4	Sequence similarity network Lipase_bact_N	260
B.5	pHAT, pHAT2, and pHAT5	260
B.6	SDS-PAGE of protein purification	261
C.1	Vector Map pET32_Strep_ACP	266
C.2	Vector Map pET32_HG3.17	266
C.3	Expression of N20 and HG3.17	267
C.4	Effects of additives on 5NBI background rate	268
C.5	The use of tartrazine to offset the signal	269
C.6	BSA activity in droplets	269
C.7	Sequencing of naive epPCR library of HG3.17	270
C.8	Sequencing of naive deletion library of HG3.17	271
C.9	Sequencing of naive insertion library of HG3.17	272
D.1	¹³ C-NMR of 6c	273
D.2	DEPT-NMR of 6c	274
D.3	HSQC-NMR of 6c	275
D.4	HMBC-NMR of 6c	276
D.5	Evidence for Homo-FRET of 6c	277
D.6	Gel-electrophoresis of the SCV plasmid library	278
D.7	Biofilm formation on chip	279

List of tables

1.1	Published Droplet Workflows for Enzyme Library Selections	17
2.1	Sorting Parameters	28
2.2	Detector Parameters	36
2.3	Projects performed on the FADS	40
3.1	Controls used to establish the esterase screen	53
3.2	Summary of the two SCV library sorting campaigns.	61
3.3	List of esterase hits	63
3.4	Characterised enzymes with homology to the hits	66
3.5	Protein family assignment of the hits	69
4.1	Reaction rate of 2a in different buffers	100
4.2	Droplet conditions tested for leakage	105
4.3	Enrichment of HG2.17 with 2a	111
4.4	Substitution libraries generated by epPCR.	119
4.5	HG3.17 epPCR library enrichment	122
4.6	HG3.17 epPCR library enrichment	127
4.7	Mutations found in the six most active variants in the cell lysate re-screening assay.	128
4.8	Michaelis-Menten kinetic parameters of the variants selected in the cell lysate assay which had amino-acid mutations for substrate 2a	130
4.9	Initial rates in cell lysate before and after sorting InDel libraries	135
4.10	Michaelis-Menten kinetic parameters of deletion variant 4A09 (DelA181, S182G) for substrate 2a	137
4.11	Summary of mutations that were found more than once and in different libraries.	142
5.1	Emission maxima in different solvents	172

5.2	Summary of the SCV library sort using 6a	174
5.3	Analysis of SCV sub-libraries	179
5.4	Sequenced variants with inserts containing complete ORFs.	183
7.1	PCR program for epPCR.	206
B.1	Composition of the SCV library	243
B.2	Blast Top Hits.	257
B.3	Cloning and expression of hits	261
B.4	Kinetic Parameter Part 1	262
B.5	Kinetic Parameter Part 2	263

Acronyms

βCD β-cyclodextrin.

λ average droplet occupancy.

AADS absorbance-activated droplet sorting.

AC class IV adenylyl cyclase.

APD avalanche photodiode.

BLAST Basic Local Alignment Search Tool.

BSA bovine serum albumin.

cfu colony forming unit.

DEP dielectrophoresis.

DMA direct memory access.

DMSO dimethylsulfoxide.

DUF domain of unknown function.

epPCR error-prone PCR.

FACS fluorescence-activated cell sorting.

FADS fluorescence-activated droplet sorting.

FIFO first in, first out.

FPGA field-programmable gate array.

FWHM full width at half maximum.

HPLC high-performance liquid chromatography.

IPTG isopropyl β -D-1-thiogalactopyranoside.

IVTT *in vitro* transcription and translation.

KSI ketosteroid isomerase.

LB Luria Bertani.

LC-MS liquid chromatography-mass spectrometry.

NaPi sodium-phosphate buffer.

NCBI National Center for Biotechnology Information.

OD optical density.

ORF open reading frame.

Oxd aldoxime dehydratase.

PDMS poly(dimethyl siloxane).

PMT photomultiplier tube.

pNP para-Nitrophenol.

PTE phosphotriesterase.

RP-HPLC reversed-phase high-performance liquid chromatography.

rRNA ribosomal RNA.

SPCM single-photon counting module.

TNA threose nucleic acid.

TTL transistor-transistor logic.

VI virtual instrument.

Chapter 1

Introduction

Enzymes drive the chemistry of life. From the original capture of inorganic compounds fundamental to life to their eventual release back into the environment, enzymes are involved in every step of the way. In this introduction, I will convince the reader that the discovery of enzymes remains an important enterprise given the sometimes surprising chemical transformations they are able to catalyse. Even for simpler chemistries, such as the hydrolysis of esters, new enzymes are increasingly sought after to help our industrial processes become more sustainable. I will explain that the screening of metagenomes, *i.e.* the screening of pooled genomes rather than a single one, has emerged as a powerful tool to discover new enzymes. However, the hit rates in functional metagenomics are low and the screening throughput limited. Thus, I will argue that combining functional metagenomics with droplet microfluidics, a method with ultrahigh-throughput, is attractive to enhance the outcomes of such screening campaigns. I will therefore describe the development of new assays for two reactions, ester hydrolysis and the Kemp elimination, to discover enzymes in the metagenome.

1.1 Enzymes

Enzymes are outstanding catalysts which drive all chemical reactions sustaining life on earth. Beyond their importance in nature, they are attractive for use in chemical synthesis given their distinct advantages over chemical catalysts: high turnover frequencies (10^7 s^{-1}), high degree of selectivity (enantio-, regio- and chemoselectivity), mild reaction conditions, and environmental friendliness [1]. Despite these advantages, the list of chemical transformations for which off-the-shelf enzymes are readily available is currently limited. The discovery of enzymes for new reactions has been constrained mostly by the failure to find them rather than the natural biodiversity: there are more enzymes to be discovered than there are known [2]. The one gene-one enzyme hypothesis formulated in the 20th century was

fundamental to the inception of molecular biology [3]. Implicit in this hypothesis was, that one enzyme is linked to one function. It is now known, that most enzymes are likely to be promiscuous, *i.e.* they catalyse several reactions and not just the chief one they evolved for, which further increases the chance of finding enzymes for a given reaction [4].

Enzyme promiscuity

Enzyme promiscuity allows chemists to use enzymes for their own purpose. In a typical application, an enzyme is employed for a reaction close to its natural activity as in the case of the ω -transaminase used in the production of the antidiabetic drug sitagliptin on an industrial scale [5]. This approach relies on the ability of an enzyme to catalyse the transformation of several substrates including the one of interest. This so-called substrate promiscuity can be used as a handle to engineer the enzyme towards enhanced activity on the synthetic substrate if needed. Enzymes also exhibit catalytic promiscuity, *i.e.* the ability to catalyse different chemical reactions, which was first conceptualised in mechanistic terms by O'Brien and Herschlag [6].

1.1.1 How to find an enzyme

Discovery of new enzyme activities has been limited by the difficulty to access the full range of natural diversity. To find a biocatalyst for a specific process, for example the making or breaking of an ester bond, four major approaches are possible: protein engineering, computational design, database mining, and metagenomics (Figure 1.1).

Engineering. If an enzyme is available for the wanted reaction, but does not meet the set performance criteria, protein engineering technologies are used to adapt or enhance enzymes for the reaction under the needed solvent, temperature, and substrate load conditions. Both rational (re)-design and directed evolution have been used to tailor enzyme properties [7, 8]. In directed evolution, a library of gene variants is created by random mutagenesis, e.g. by error-prone PCR, followed by selection of phenotypically superior mutants. The selected gene variants are then subjected to the same process, thus completing several cycles of variation, selection, and amplification [9]. The evolutionary approach requires functional expression of the enzyme and a robust screening or selection strategy but not necessarily a high degree of insight into the enzyme's mechanism of action. Detailed characterisation of the enzyme is needed in rational design and re-purposing. Even if often guided by "chemical intuition", the structure and a detailed understanding of the catalytic mechanism are needed to identify residues for mutation and potential catalytic activities to explore [10].

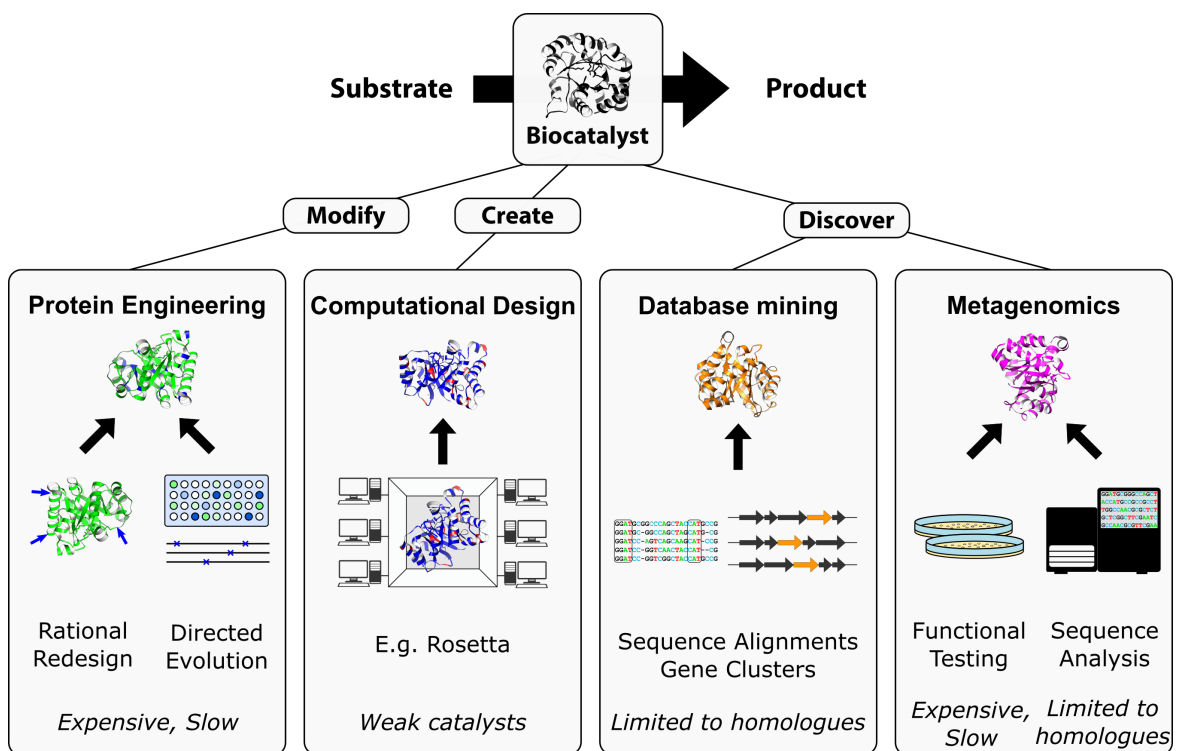


Fig. 1.1 There are four major sources for new enzymes: engineering of existing proteins, design of new catalysts, database mining, and metagenomics.

Design. If there is no biocatalyst available that can be modified, an enzyme can be created in rare cases. Enzymes have been generated by arranging amino acids around a calculated transition state and docking this arrangement into pre-existing protein scaffolds. As was the case for Kemp eliminases, retro-aldolases, and diels-alderases [11–13]. These are catalysts by design, a concept once dubbed the “Holy Grail” of chemistry by Wong and Whitesides [14]. While the crystal structure of these enzymes was close to the intended design (within 1 Å) and catalytic activity was detectable, the initial catalytic efficiencies were low (the highest was $10^3 \text{ M}^{-1} \text{ s}^{-1}$ for the Kemp eliminase HG3 [15]). The enzymes needed improvement by directed evolution to reach activities reminiscent of natural catalysts [16]. As long as this limitation persists, the screening of the natural biodiversity for functional catalysts remains an attractive alternative.

Mining. One cost-effective way to screen the natural diversity is to mine databases *in silico*. Sequence databases are growing rapidly due to massive next-generation sequencing efforts [17]. The PFAM database (PFAM 31.0, March 2017) distinguishes 16,712 protein families of which 3918 are classified as domains of unknown function (DUFs) [18, 19]. This illustrates just how much chemistry may yet be discovered with over 20% of protein families not having a known function at all. However, functional annotation in databases is based on sequence homology, which itself is inferred from similarity between the new and previously-assigned sequences [20], see also Section 1.4. Therefore, the success of database mining depends on the existence and accuracy of previous assignments. Errors in databases can be misleading. Recently, the first enzyme catalysing N-N bond formation was found by sequence analysis of biosynthetic gene clusters. Its sequence had been mis-annotated as a transcriptional regulator with a flavin binding domain in the database. The enzyme, found to use a heme cofactor, was not a regulator and catalysed this unusual reaction [21, 22]. As in this study by Du *et al.*, database mining can be guided by natural compounds containing an unusual chemical moiety and their associated gene clusters. Identifying homologous, uncharacterised genes that appear in all the clusters and testing them may yield enzymes with truly novel catalytic properties. However, such an approach requires extensive knowledge of natural compounds and their respective gene clusters.

Metagenomics. The only way to avoid pre-selective assumptions regarding either the sequence, structure, or mechanism of the desired catalyst would be by performing a direct biochemical test on a library of expressed proteins. Accessing the natural biodiversity in this way has become possible thanks to functional metagenomics and will be discussed in more detail in the following section.

1.2 Metagenomics

It was long noted in microbiology that only a fraction of the microorganisms from any environment could be cultured under laboratory conditions. The effect was known as “*the great plate count anomaly*”, referring to the fact that the number of colonies growing on a culture plate was, in the best of cases, only 1% of the number of cells observed under the microscope [23]. This led to the development of culture-independent methods to study the diversity of microorganisms based on the direct sequencing of genetic material isolated from an environment.

1.2.1 Early studies and metagenomic sequencing

Inspired by the groundbreaking work of Carl Woese on ribosomal RNA (rRNA) as a phylogenetic marker, early studies on microbial diversity focused on the genes encoding 16S rRNA [24, 25]. These efforts led to the creation of the first recombinant DNA library containing genetic material from a mixed population of organisms rather than a single one [26]. In the latter study, 15 unique rRNA sequences were found in a library from the Sargasso Sea of which 2 were from previously unknown phyllogenetic groups.

The development of shot-gun next generation sequencing revolutionised the field. In 2004, a single such study (again of the Sargasso Sea) found at least 1800 genomic species with 148 constituting new phylotypes [27]. Importantly, the power of next generation sequencing allowed the shift towards the analysis of existing genes and their frequencies in a given environment. It also opened the door to the discovery of novel genes. The 2004 Sargasso Sea study alone reported over one million previously unknown genes [27]. To date, almost 34,000 projects and near to 800,000 samples associated with metagenomic sequencing were deposited on the BioProject and Biosamples database of National Center for Biotechnology Information (NCBI).¹ Most of the projects are very recent: 80% of the 50,000 metagenomic sequencing projects on the manually-curated genome online database (GOLD) were deposited while this thesis was carried out, *i.e.* in the last three years [28].

These sequencing projects yielded some notable discoveries. Beja *et al.* identified the proteorhodopsins of marine bacteria [29]. These proteins are homologues of bacteriorhodopsin, a light-driven proton-pump which was previously thought to only exist in archaea. Similarly, ammonium monooxygenases, which were previously thought to only exist in bacteria, were discovered in archaea [30]. Both findings had a major impact on their respective fields of research. However, they were also only possible due to the existence of homologous sequences associated with a known, *i.e.* experimentally verified, function. To

¹As of September 2018. The search term used was “metagenom*”.

discover genuinely novel enzymes, approaches solely based on sequencing are unsuitable *a priori*. By reviewing the literature, Ferrer *et al.* found that while only 200 new enzymes were discovered and characterised using sequencing based methods, almost 6,000 had been isolated using a functional metagenomic approach [31].

1.2.2 Functional metagenomic screening

Metagenomic libraries are constructed in analogy to classical genomic libraries. A general workflow is shown in Figure 1.2. The DNA of interest is extracted and purified from an environmental sample. It is then physically fragmented or enzymatically digested and ligated to a vector for transformation into a host organism [32]. The constructed library is then subjected to a functional assay to detect hits, *i.e.* clones that exhibit the desired activity. The most frequent type of assay is a phenotypic screen, for example the formation of a clear zone around a colony on a skimmed-milk plate for protease activity [33]. Another approach is heterologous complementation, where the host lacks a function which is essential for survival or detectable growth which can then be complemented by the recombinant DNA [33]. Healy *et al.* reported the first functional metagenomic screen [34]. They isolated DNA from an anaerobic digester, cloned it into the vector pUC19, transformed it into *E. coli*, and screened 15,000 colonies using a plate assay for cellulase activity. They detected 23 clones, but reported only one new sequence. This study is still representative of the typical hit rate of a functional metagenomic screen: 10^{-5} to 10^{-4} [31]. Besides the chosen environment and quality of the extracted DNA, the number of hits obtained from a screening is affected by three major factors: the library format (*i.e.* the chosen vector), the efficiency of heterologous expression, and the throughput of the detection assay.

Vector system. Libraries fall into two principal categories: small insert libraries up to 10 kbp in plasmids and large insert libraries at 40 kbp in fosmids or cosmids (or even larger inserts in bacterial artificial chromosomes) [35]. Smaller inserts are more likely to be transcribed effectively due to higher copy numbers and possible transcription from the promoter of the vector. This helps to detect weakly expressed metagenomic genes [35, 36]. Larger insert libraries allow the detection of large genes and gene clusters, a smaller number of clones is required to screen a set number of genes, and the insert size is more consistent thus reducing cloning biases [37]. However, they rely more on transcription from within the DNA insert rather than the vector and have lower copy numbers.

Expression host. As implied above, either library format relies on the host organism's ability to transcribe foreign DNA. Transcription was shown to be enhanced by randomly in-

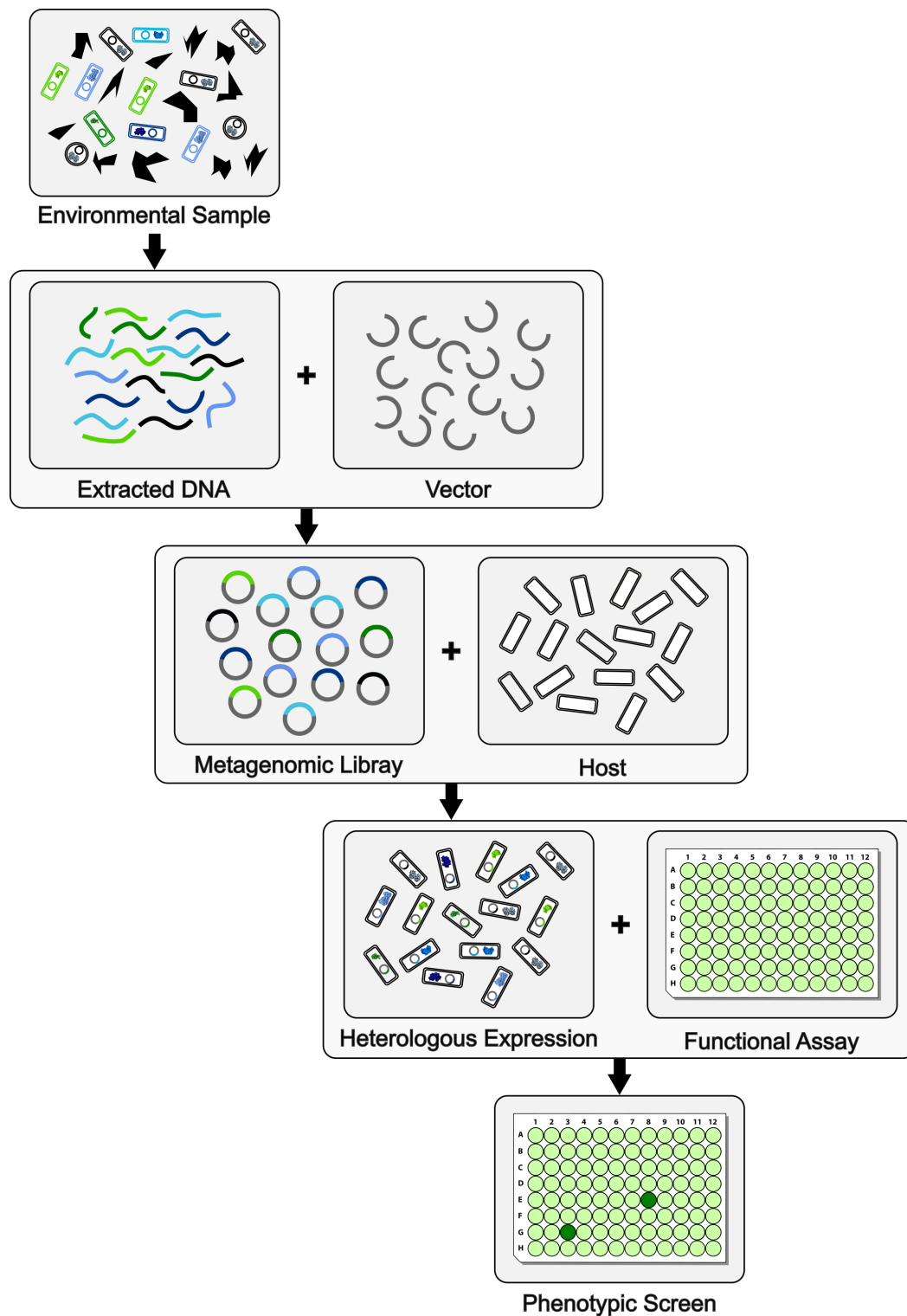


Fig. 1.2 A general workflow for the functional screening of a metagenomic library. Once the metagenomic DNA is extracted, there are three major decisions to make that influence the success of a screening campaign: which vector to use to construct the library, which host organism to express the library in, and which assay to use to detect the desired enzymatic activity.

roducing promoter sequences into the metagenomic DNA [38, 39], by using bidirectional transcription [40], or by introducing heterologous σ -factors into *E. coli* [41]. However, transcription is only the first step in the expression of a functional protein. The translation machinery, the absence of required post-translational factors such as chaperones, and toxicity to the chosen host may be limiting access to new enzymes [33, 42]. The use of different hosts such as *B. subtilis* can lead to the identification of different hits [43], or improved hit rates if the aim is specific, such as finding thermostable enzymes using *T. thermophilus* as a host [44]. An entirely new class of enzymes was discovered using heterologous expression in cyanobacteria: the enzymes involved in the synthesis of poly-brominated aromatic compounds [45]. However, *E. coli* remains the host of choice in most studies to date, because of its high transformation efficiency and ease of use.

Assay throughput. The third factor affecting how many hits are obtained is the throughput of an assay, which is the main limitation addressed in this thesis. Tallying the numbers reported by Simon and Daniel, the median number of clones tested in phenotypic screens for fosmid/cosmid libraries is less than 10^4 and for plasmid based libraries less than 10^5 [33]. Usually, less than ten positive clones are found resulting in the typical hit rates of 10^{-5} to 10^{-4} [31]. It stands to reason, that if the throughput of the phenotypic screens were increased by a factor of ten, ten times more hits would be obtained. The technology which allows such a boost in throughput is droplet microfluidics, the topic of the next section.

1.3 Droplet microfluidics

Droplet microfluidics provides a screening technology which allows us to routinely assay 10^7 library entities per day, thus providing ultrahigh-throughput [46]. Any enzyme discovery and enzyme engineering/evolution campaign benefits from such increased screening capability compared to traditional methods, as shown for example by an improved outcome in Obexer *et al.* compared to Giger *et al.* [47, 48]. Hence, droplet microfluidics is becoming an increasingly important technology. In the following paragraphs, the central concepts in library screening using droplets will be laid out.

In droplet microfluidics the biological entities, *e.g.* *E. coli* cells, are compartmentalised into monodisperse water-in-oil droplets. The reaction vessel in which the bioconversion takes place is the water droplet (typically 2 to 200 pl in volume) which is spatially separated from all other reaction compartments by the oil phase. There must be no, or very limited, exchange of contents between the droplets to ensure genotype (DNA sequence) and phenotype (enzyme activity) linkage just as in living cells (Figure 1.3) [49, 50]. Thanks to the

miniaturisation to picolitre volumes, only small amounts of valuable chemicals are required to perform a large number of reactions.

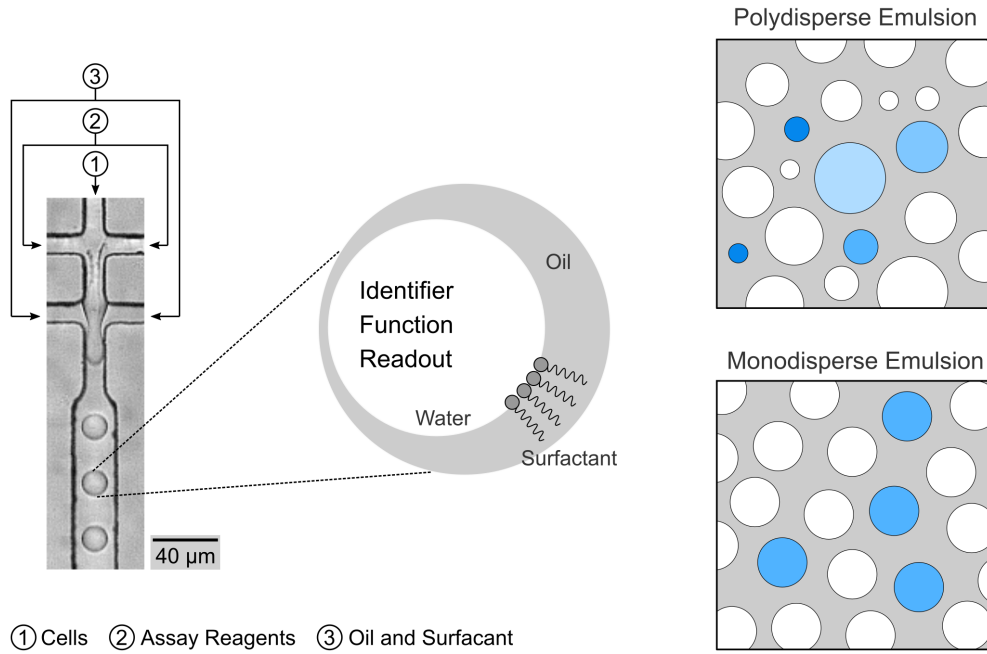


Fig. 1.3 *Left*: Shown is droplet generation in a microfluidic flow-focusing device. In a typical application, cells and assay reagents are mixed on chip, before the aqueous phase is sheared off in oil containing a surfactant. The droplet boundary ensures that the identifier (gene) of the function (enzyme activity) and its readout (e.g. fluorescence) remain linked. Adapted from [50]. *Right*: In polydisperse emulsions the cytosol of cells is diluted to different degrees, *i.e.* therefore the same genotype yields different levels of signals. In monodisperse emulsions, the same signal is obtained for the same genotype.

One factor limiting the throughput of library screening in droplets is the probabilistic nature of cell encapsulation (and of other library entities such as plasmids). The distribution of cells in solution and therefore their arrival at the encapsulation site is random [51]. At high cell concentrations the number of cells per droplet follows a normal distribution. However, in library screening it is desired to have only one cell present in one droplet and thus lower concentrations need to be used. The number of cells contained in a given droplet is then governed by the Poisson distribution [51]:

$$P(\lambda, k) = \frac{\lambda^k e^{-\lambda}}{k!}; \quad \lambda = \frac{[\text{Cells}]}{\bar{V}_{\text{droplet}}} \quad (1.1)$$

With λ being the average number of cells per droplet (occupancy), k the number of cells per droplet, $[\text{Cells}]$ the concentration of cells, and \bar{V}_{droplet} the mean droplet volume. The highest

achievable proportion of droplets containing one cell only is 37% at a λ of 1. However, as shown in Figure 1.4B, the proportion of droplets with 2 or more cells would be 8%. Co-encapsulation of several cells can severely limit the ability to enrich improved library variants. Therefore, many reported studies use a λ of 0.1 to limit co-encapsulation, which means that 90% of droplets are empty. However, the ability to generate and manipulate tens of millions of droplets means that the throughput of droplet microfluidics still exceeds traditional screening methods by several orders of magnitude.

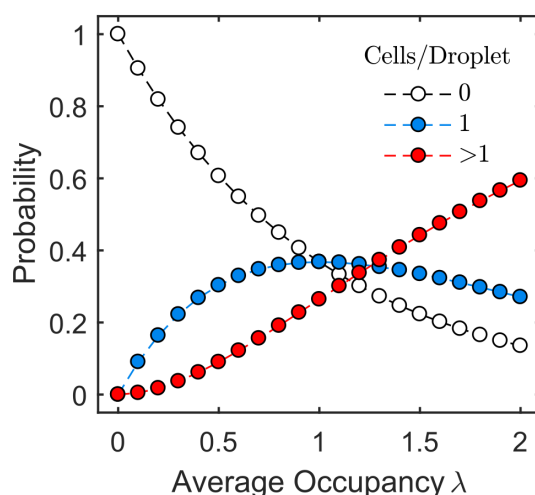


Fig. 1.4 Likelihood that a droplet contains 0, 1, or more than 1 cells depending on the average occupancy λ of the Poisson distribution.

In its earliest form, the generation of picolitre droplets for a library selection was achieved by simply mixing oil, surfactant, and an aqueous solution using a magnetic stirrer bar. Thus it was possible to enrich the HaeIII methyltransferase gene out of an excess of 10^7 dihydrofolate reductase genes [52]. Although technically simple to generate, the droplet size distribution in such emulsions is very sensitive to the mixing conditions limiting the reproducibility of experiments [53]. Furthermore, the size distribution of such droplets is generally wide with the standard deviation often larger than the mean diameter, *i.e.* they are polydisperse. This introduces bias limiting quantitative comparison between the droplets in screening campaigns (Figure 1.3) [53, 54]. Monodisperse droplet populations with only a few percent variation around the mean diameter overcome these limitations and are generated and manipulated using microfluidic devices.

The fabrication of microfluidic devices has become widely accessible thanks to the development of soft lithography [55]. The general steps are outlined in Figure 1.5. In soft lithography, replica molding is used to create a micropatterned elastomer. The liquid prepolymer of the elastomer is poured over a master with a relief structure on its surface, allowed to

polymerise, and then peeled off. The most commonly used elastomer is poly(dimethyl siloxane) (PDMS). The master is obtained by photolithography. In photolithography, a pattern is transferred from a photomask to a light-sensitive chemical, the photoresist, on a substrate. The substrate usually is a single-crystal silicon wafer and the most widely used photoresist for soft lithography is SU-8 [56]. SU-8 is a negative photoresist, *i.e.* where exposed to UV light it polymerises. Unpolymerised SU-8 is removed by washing with a solvent, revealing the intended microfluidic channels as a relief structure [56]. Several dozen of PDMS replicas can be created from a single silicon/SU-8 master. A replica is bonded to a glass or second PDMS surface after treatment with an oxygen plasma to seal the imprinted microfluidic channels. The plasma-activated channel surfaces can be modified with a trichloro-silane derivative to obtain the required surface properties [55].

In 2001, the first monodisperse droplets using a PDMS microfluidic device were generated [57]. Since then, a plethora of geometries and devices have been designed that can be used as individual modules to re-create macroscale laboratory procedures on the microscale [58, 50]. It is now possible to generate highly monodisperse droplets at frequencies of several kilohertz, to incubate them, to inject solutions into them, as well as to perform quantitative measurements like the selection of droplets according to their fluorescence (fluorescence-activated droplet sorting, FADS) or absorbance (absorbance-activated droplet sorting, AADS) [59–61]. Moreover, the transformation of microfluidic emulsions to double emulsions (water-in-oil-in-water compartments) or hydrogel beads equipped with polyelectrolyte shells are compatible with fluorescence-activated cell sorting (FACS) [62, 63]. So far, droplet-based microfluidics have been successfully used for directed evolution, strain selection, bioprospecting, and in one case the screening of a metagenomic library which will be discussed below [64–67].

1.3.1 Screening enzyme libraries using droplet microfluidics

The basic workflow to screen enzyme libraries in microfluidic droplets is illustrated in Figure 1.6. A gene library is first transformed and expressed by a host organism (1). Single cells are encapsulated together with a chosen substrate and a lysis agent into water-in-oil droplets (2). Each cell lyses inside its droplet causing the release of the expressed enzymes (3). Crucially, the droplet boundary ensures retention of both genotype and phenotype. Depending on the activity of a given enzyme variant, a substrate is turned into product. Droplets above a set amount of product are collected using droplet sorting (4). After sorting, the genetic material can be recovered and the phenotypically superior mutants expressed for further characterization or for further enrichment (5).

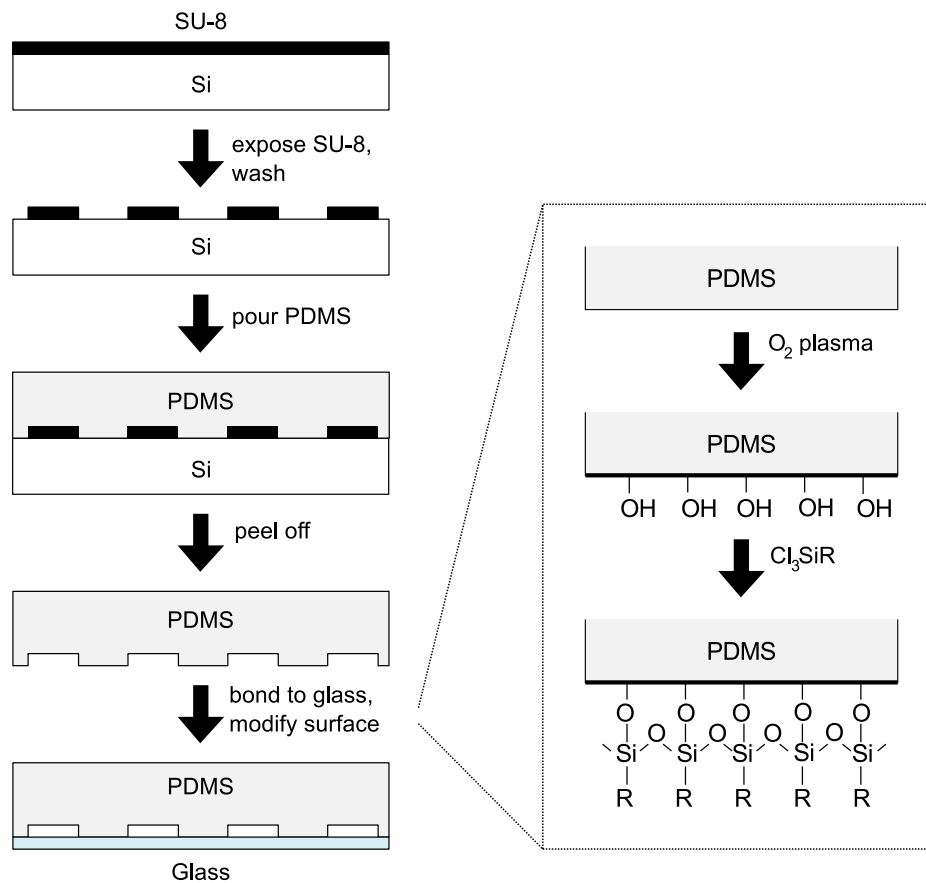


Fig. 1.5 The steps in the fabrication of a microfluidic PDMS device. First, a silicon substrate is patterned with the photoresist SU-8 to generate a master. Using the master, many PDMS devices can be generated via replica molding. The patterned PDMS is plasma-bonded to another PDMS surface or a glass substrate to seal the microfluidic channels. The activated PDMS/glass surface is usually modified using trichloro-silane derivatives, *e.g.* to create a fluorophilic layer to allow the use of fluoruous oils. Adapted from [55].

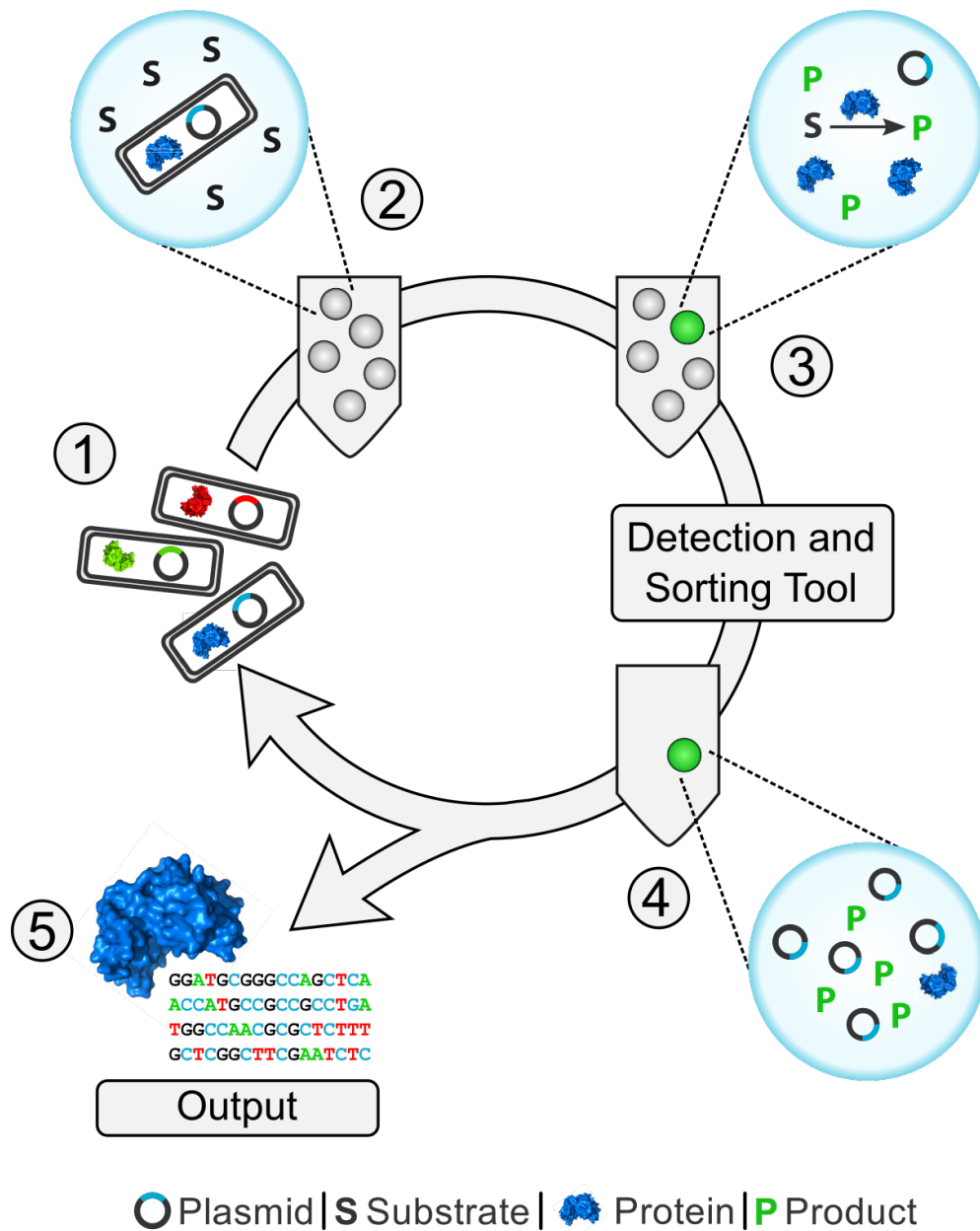


Fig. 1.6 A typical workflow for the screening of an enzyme library using microfluidic droplets. *1:* A gene library is first transformed and expressed by a host organism. *2:* Individual cells are encapsulated into picoliter water-in-oil droplets together with a chosen substrate and lysis reagents. *3:* After cell lysis, the expressed enzymes are released and, if active, convert the substrate into product. *4:* If the fluorescence of a droplet exceeds a set threshold it is physically sorted using a bespoke microfluidic device. *5:* After sorting, the genetic material can be recovered and the phenotypically superior mutants expressed for further characterization. Adapted from [46].

After a decade dominated by proof-of-principle experiments in droplets, the utility of this new technology was shown in several library screening campaigns [61, 66–69]. The first directed evolution campaign in microfluidic droplets was targeting horseradish peroxidase, using a coupled assay and lead to over 10-fold improved catalysts [64]. In addition to such highly active enzymes, even slow reactions can be screened, as exemplified for a sulfatase: the thermodynamically most challenging biochemical reaction is amenable to the droplet format [67, 69]. These experiments demonstrate the range of possible timescales with incubations between 5 minutes and 48 hours, suggesting that fast and slow reactions can be monitored.

1.3.2 Enzyme substrates in droplet experiments

For droplet sorting to be successful, the used substrates and their products play a critical role. The availability of suitable substrates is what currently limits the range of enzymatic assays amenable to droplet microfluidic methods. The chief requirement is that the reaction product and any long-lived intermediates do not exchange between the droplets otherwise the geno- to phenotype linkage is lost [49].

A number of established fluorescent or absorbing leaving groups used in enzyme assays do readily exchange between droplets, for example rhodamine, coumarin derivatives, and 4-nitrophenol [70]. This observed *leakage* of product molecules may be due to direct transfer through the carrier oil or by micellar transport [71, 72]. Exchange can be reduced by using fluorinated carrier oils and by using additives such as bovine serum albumin (BSA) [73].

An effective way to prevent leakage is the chemical modification of the leaving group. Most studies aimed at installing charged groups on the substrate to reduce solubility in the oil phase. Several authors succeeded in eliminating substrate/product leakage by introducing a sulfonic acid, which due to its low pK_a introduces an effectively permanent negative charge. Najah *et al.* introduced a sulfonic acid group into a coumarin derivative; no exchange between droplets was detected over 24 h at 30 °C. The resulting substrate was used to screen for cellobiohydrolase activity on model bacterial strains [74]. Similarly, Fenneteau *et al.* generated a sulfonylated rhodamine-based enzymatic substrate and used it to assay amino-peptidases [75]. A permanent positive charge was used to improve a previously optimized artificial retro-aldolase in droplets by introducing a charged ammonium group into a methodol derivative [47]. Ma *et al.* followed a similar approach for phosphotriesterase substrates with coumarin derivatives as leaving groups [76]. This study specifically exploited the differential solubility of substrate and product, so as to trap only the product inside droplets, but allow the substrate to diffuse in and out of droplets. This enabled the quick addition and removal of substrate to all droplets at defined times, increasing the control over the reaction time.

1.3.3 Enzyme assays available in microfluidic droplets

Thanks to the development of substrates and droplet workflows the range of enzymatic assays available in droplet format is increasing. An overview of the assays currently available are reviewed below and the corresponding workflows are shown schematically in Table 1.1.

Hydrolases. Hydrolytic reactions in which water displaces a fluorescent leaving group are perhaps the reactions that can be assayed most simply, as the reaction product itself is detectable. Directed evolution of triesterases, phosphonate hydrolases, and enrichment experiments for glycosidases as well as sulfatases were successful [62, 63, 69, 77]. Screening of natural gene repertoires for glycosidases, amylases, and triesterases were demonstrated [66, 67, 78].

Aldolases. Obexer *et al.* accomplished the directed evolution of a computationally-designed aldolase which had previously been evolved in a microtiter-plate screening assay [48, 79]. Notably, in this study the enzyme evolved a strong preference for the (S)-enantiomer of the substrate. In a similar study, the same authors achieved (R)-selectivity by choosing an evolutionary trajectory which had not been accessible in the classic screening format [47].

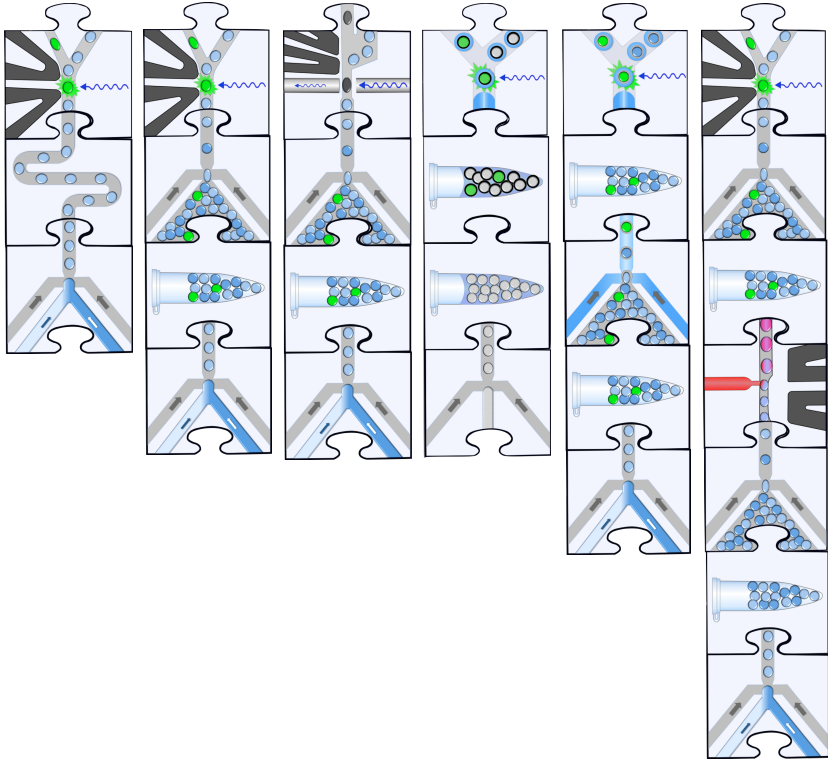
Polymerases. The above studies monitored formation of a fluorescent product directly. In a different approach, Larsen *et al.* used a molecular beacon to evolve a non-natural threose nucleic acid (TNA) polymerase [80]. This assay was performed in a double emulsion, which is compatible with commercial FACS devices circumventing the need for a custom-built fluorescence-activated droplet sorting (FADS) setup [62]. Using a molecular beacon and the multistep format, Ryckelynck *et al.* achieved the implementation of a full SELEX workflow in order to evolve a ribozyme [81].

Proteases. Although no library screening was achieved, protease activity has been successfully detected in droplets. A single-cell secretion assay in droplets employing multiple protease substrates and multi-colour analysis was implemented based on fluorescence quenchers [82]. Price and Paegel developed a droplet assay to detect inhibition of HIV-1 protease activity by UV-induced release of defined amounts pepstatin A from beads [83].

Amino-acid dehydrogenases. A coupled reaction was implemented to select improved L-phenylalanine dehydrogenases. This study established the first AADS module. The coupling of cofactor reduction with production of a tetrazolium dye enabled a 25-fold signal amplification for product detection compared to direct detection of NADH. Similar assays are

available for dehydrogenases and reductases and should make these reactions amenable to screening in droplets [61].

Table 1.1 Tabular overview of recently published enzyme library selections. The jigsaw pieces represent steps in microfluidic workflows that can be recombined as needed (as in [58]).

Workflow		Sorting	Enzyme	Reference
		FADS	aldolase	[47, 79]
		FADS	α -amylase, aldolase, phospho- triesterase, phosphonate hydrolase, sulfatase	[47, 62, 63] [67, 77, 69] [79]
		AADS	amino acid dehydratase	[61]
		FACS	sulfatase	[63, 67]
		FADS	TNA polymerase	[80]
		FADS	X-motif ribozyme	[81]

1.4 Ultrahigh-throughput metagenomics in droplets

As laid out in the previous sections, functional metagenomics is a powerful method to find new enzymes, albeit with low success rates and limited throughput. Droplet microfluidics is a technology that allows the screening of biological entities at ultrahigh-throughput. Using droplet microfluidics to enhance the outcomes of functional metagenomics is thus an attractive enterprise. There has been one previous study which achieved this.

Colin *et al.* screened a plasmid-based library, the SCV library, consisting of more than one million members for sulfatase and phosphotriesterase activity, see Appendix Table B.1 [67, 84]. Remarkably, it was possible to detect enzymatic activity using a single-cell lysate assay as outlined in Figure 1.6. Six unique DNA inserts encoding new sulfatases and eight inserts encoding new phosphotriesterases were isolated using FADS. The phosphotriesterases would have been particularly difficult to isolate using traditional methods: there is no agar plate assay available, only much more time- and resource-intensive well plate assays. The average catalytic efficiency (k_{cat}/K_m) on the substrate used for selection was $10^3 \text{ M}^{-1} \text{ sec}^{-1}$. Such low activities would have further limited the ability to reliably identify them from a library of millions in a classical approach. Only one of the open reading frames (ORFs) showed homology with a previously characterised phosphotriesterase. None of the other hits would have been identifiable using a sequence-based approach. Extending and applying the microfluidic screening platform pioneered by Colin *et al.* to more reactions is likely to yield more catalysts, but may also yield fundamental insights not otherwise possible. The conditions for the success of the ideas outlined below are that the assays are sensitive, that hit confirmation is simple, and that sequence analysis and ORF assignment are reasonably fast.

Enzyme annotation. Currently, the annotation of new sequences is based on sequence similarity searches against databases. If the similarity is higher than would be expected for two random sequences, this implies homology, *i.e.* that the two sequences are related by common ancestry. Evidence for homology is then further extrapolated to the gene products having a similar function [20]. This line of thinking has several limitations:

- With the increase in database sizes the chance of randomly detected homology has increased, leading to potentially false annotations from the outset [85].
- While such false positives are limited for the most commonly used algorithms, there is no assessment of their false negative rates [20]. Structure is more conserved than sequence, implying that while homology may clearly exist in terms of structure, it may not be detectable at the sequence level [86]. Also, some enzyme classes, such as the hydroxynitrile lyases, show little homology due to convergent evolution [87].

- The relationship between homology and function is complex, in particular for enzymes where very few amino acid substitutions can change or abolish the catalytic activity [20, 88].
- The vast amount of annotated sequences available on the databases today are from whole genome or metagenomic sequencing projects. Therefore, the annotation of large clusters of sequences may only be underpinned by very limited functional evidence.

The ultrahigh-throughput functional tests possible in droplets provide data that complement and correct sequence analysis, allowing a better understanding of the existing sequences. As mentioned above, the activities of the phosphotriesterase hits identified by Colin *et al.* suggested by sequence similarity could rarely be validated, while the droplet screening established a functional assignment [67, 84]. Functional metagenomics in droplets thus provides a means for more comprehensive annotation based on experiments.

New bridgeheads for functional annotation. Sometimes enzymatic activities are particularly difficult to predict, for example when very few experimentally-characterised examples exist. Given the difficulty in predicting primary activities, the prediction of promiscuous reactions of enzymes is even more challenging. These can be understood as evolutionary starting points: after gene duplication they can be further refined by adaptive evolution and lead to re-functionalised proteins [6, 89]. There is increasing experimental evidence, that promiscuous activities play an important role in the adaptation of metabolic pathways [90, 91] The more such activities are known, the more can they be understood in biology and be exploited for industrial applications. The activities identified by Colin *et al.* were largely such promiscuous activities [67, 84].

Simulating early evolution. A scenario in which an expressed metagenomic library is tested against a non-natural substrate mimics the encounter of environmental microbiota with compounds introduced by mankind (as *e.g.* in the case of phosphate triesters). Evidence suggests that an initial response to novel compounds already exists even in naïve environments [6, 89]. These starting points based on promiscuous enzymes then further evolve to create efficient pathways that extract energy and nutrients [91, 92]. A droplet microfluidic assay may provide the throughput needed to identify these rare starting points for many reactions. Performing rounds of directed evolution following the identification of a promiscuous enzyme could lead to improvement of the activity of the target reaction. Together, such an experiment would constitute a simulation of the early evolution of a new function.

1.4.1 Goals of this thesis

To investigate the three hypotheses outlined above, three major goals were set in this thesis:

- **Building a fluorescence-activated droplet sorter.**

At the start of this work no FADS was available. Given that such a set-up was the pre-condition to setting up many types of assays, a new sorter was built and tested. It was then used successfully in several collaborative projects as described in Chapter 2.

- **Establishing a metagenomic droplet assay for esterases.**

Interestingly, no major study in droplets has reported on the directed evolution or selection of esterases. Yet, they are most frequently screened for in functional metagenomics [93]. Esterases are a well-studied class of enzymes, common in nature, and of industrial interest. A droplet microfluidic assay to screen for such fundamentally important enzymes was deemed of high interest. Therefore, a functional metagenomic droplet screen for esterases was developed in Chapter 3

- **Establishing a metagenomic droplet assay for Kemp eliminase.**

The Kemp elimination is the general base catalysed rearrangement of 1,2-benzisoxazole to cyanophenol [94, 95]. It is a reaction not known to naturally occur in living organisms which has been used extensively to engineer chemical and biological catalysts, as will be reviewed in Chapter 4. This reaction requires only a general base within a hydrophobic environment to achieve high catalytic rates and is therefore a good starting point to screen a metagenome for promiscuous enzymes.

To establish and perform these assays, the goal was to follow the approach outlined in Figure 1.7. First, the sorting technology had to be established, *i.e.* the optics, electronics, and data analysis had to be set up. Next, the assay conditions needed to be tested using a positive and negative control. Conditions to distinguish the controls were established in well plate assays and then tested in droplets. Once suitable conditions were found, they were validated by enriching the positive over the negative control. This finally allowed the screening of the metagenomic library (the SCV library by Colin *et al.* was used). After droplet screening, a secondary screen was required to retrieve hits. The Kemp eliminase screen is currently at this step. In the esterase screen, it was possible to retrieve the twelve unique hits and sequence the DNA inserts. The most likely ORFs to encode for an esterase were cloned and the esterase activity of thirteen genes was confirmed. After this, the hits were expressed, purified, and characterised kinetically and in terms of their promiscuity.

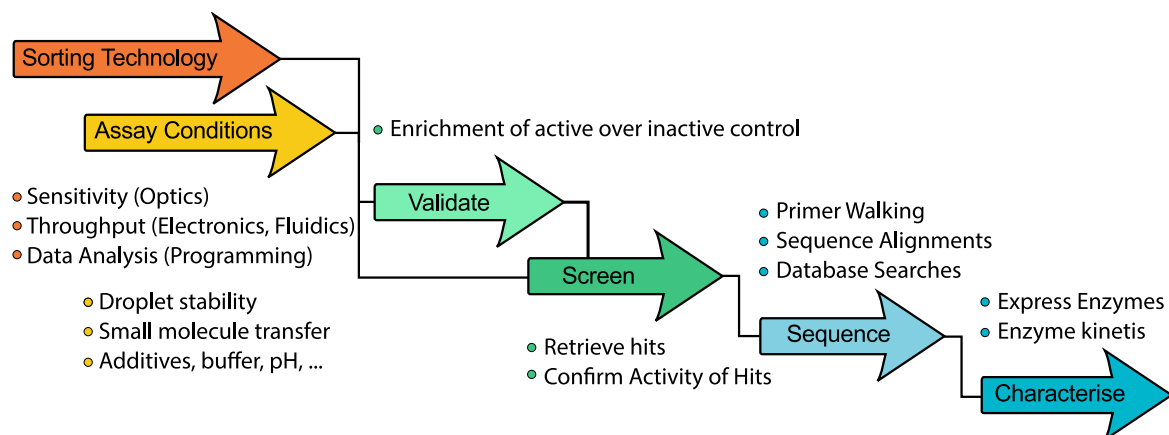


Fig. 1.7 General steps necessary to establish a microfluidic droplet screening system for metagenomics.

Chapter 2

The fluorescence-activated droplet sorter

2.1 Abstract

The aim of this chapter is to present the fluorescence-activated droplet sorter (FADS), which was built over the course of this project. The sorting principle in FADS is analogous both in name and principle to FACS. Aqueous droplets pass a laser beam and those droplets which meet the sorting criteria are deviated from their trajectory by an electric field for collection. In FACS the droplets are generated in air, while in FADS they are suspended in oil. A FADS system can be divided into a sorting chip, which acts as a disposable flow cell, and the instrumentation to control the sorting process. The focus of the work presented here was to build and incrementally improve the latter. I will now briefly introduce the key concepts in droplet sorting and follow with a detailed description of the instrument. I will illustrate how the system was improved based on the needs of key experimental procedures used in the laboratory and conclude with possible future improvements to the system.

Contributions: *All of the work presented in this chapter is entirely my own. I built the optical instrument, developed the control electronics and programmed the hardware. I wrote the code for real-time acquisition and display of the data as well as the post-sorting scripts to analyse the complete datasets. I additionally contributed to a series of collaborative projects as mentioned in Table 2.3. In the sulfatase project with Dr Bert van Loo I performed all of the microfluidic work: I optimised the reaction conditions in droplets, performed enrichment experiments using controls and sorted the mutant libraries. In the protease project with Dr Josephin Holstein I performed all of the droplet sorting and adapted the real-time signal processing on my instrument without which it would have been impossible to isolate improved enzyme variants from the six mutant libraries I sorted. I trained and performed initial droplet sorting with David Schnettler and Stefanie Neun, after which they were able to use the instrument independently.*

2.2 Introduction

The creation of large enzyme libraries is routine in molecular biology. Yet, their comprehensive functional assessment is not. For example, metagenomic libraries with up to 10^7 members have been reported [37]. Similarly, random mutagenesis by error-prone PCR creates so many variants, that the library size is usually limited by the number of transforming cells (up to 10^{10} cfu/ μ g for *E. coli*, [96]). This diversity outstrips the screening capability of traditional assays on culture or in well plates by at least two orders of magnitude. In consequence, libraries are severely under-sampled. With beneficial mutations for a desired activity being rare, the throughput becomes a limiting factor for the success of a library screening campaign [9].

Droplet microfluidics has emerged as one technology providing the ultrahigh-throughputs needed to address this issue. Conceptually, microfluidic water-in-oil droplets are miniaturised reaction vessels that bring biochemical tests from the microlitre to the picolitre range. Using a flow-focusing geometry, it is possible to generate picolitre-volume droplets at frequencies of over 10 kHz with excellent monodispersity [97, 98]. At such frequencies over 10^7 droplets, *i.e.* independent biochemical reactions, are created in one hour. As in classical screening approaches, the reaction of interest can be linked to an optical readout such as fluorescence. However, as opposed to culture and well plate techniques, the content of each droplet cannot be accessed freely. It is therefore necessary to isolate droplets with the desired level of fluorescence *via* droplet sorting.

A number of active droplet sorting techniques have been developed using pneumatic, thermal, magnetic, electric, and acoustic forces to actuate the droplets [99]. The highest throughputs (routinely 2-3 kHz) can be achieved by electric and acoustic droplet sorting [99]. While acoustic sorting has been successfully demonstrated, the required instrumentation and chip fabrication is complex and library selections remain to be demonstrated [100, 101]. Using an electric force to actuate droplets, the simplest chip design for sorting requires only a Y-junction to separate positive from negative droplets and a pair of electrodes as shown in Figure 2.1. The introduction of salt water electrodes simplified the fabrication of chips for FADS markedly, eliminating the need for re-using sorting chips [102].

The most widely used sorting principle to date is based on dielectrophoresis (DEP). The possibility to use DEP for droplet sorting was first investigated by Ahn *et al.* and I will here follow their analysis of the physical principles in action, which will help to understand the influence of different sorting parameters, *e.g.* droplet size and oil viscosity [103]. DEP exerts a force on a particle in a non-uniform electric field. It does not require the particle to have a net charge, but is caused by an induced dipole moment which drives the movement of the particle along or against the electric field gradient [104]. The force acting on a particle in its

medium, in this case a water droplet in oil, is given by [103, 104]:

$$\vec{F}_{\text{DEP}} = (\vec{m} \cdot \nabla) \vec{E}_{\text{rms}} \quad (2.1)$$

Where \vec{m} is the dipole moment of the droplet and \vec{E}_{rms} is the root-mean-square of the applied electric field. The dipole moment is given by [103, 104]:

$$\vec{m} = 4\pi\epsilon_{\text{oil}}\text{Re}[\text{CM}(\omega)]r^3\vec{E}_{\text{rms}} \quad (2.2)$$

With ϵ_{oil} being the relative permittivity of the oil phase, $\text{Re}[\text{CM}(\omega)]$ being the real term of the Clausius-Mosotti factor, and r being the radius of the droplet. The sign of the Clausius-Mosotti factor determines if the droplet is attracted to (positive DEP) or repelled by (negative DEP) areas with a high electric field gradient. This factor takes into account the conductivity and permittivity of the oil phase and droplet as well as the frequency of the electric field [104]. In principle, DEP is possible in DC fields (frequency = 0). In practice, AC fields (frequency > 0) are used to avoid heating and electrolysis at the electrodes [105]. For water droplets in oil, the real part of the Clausius-Mosotti factor is predicted to be +1 (positive DEP) over a large range of frequencies, which Ahn *et al.* confirmed in their experiments [103]. The

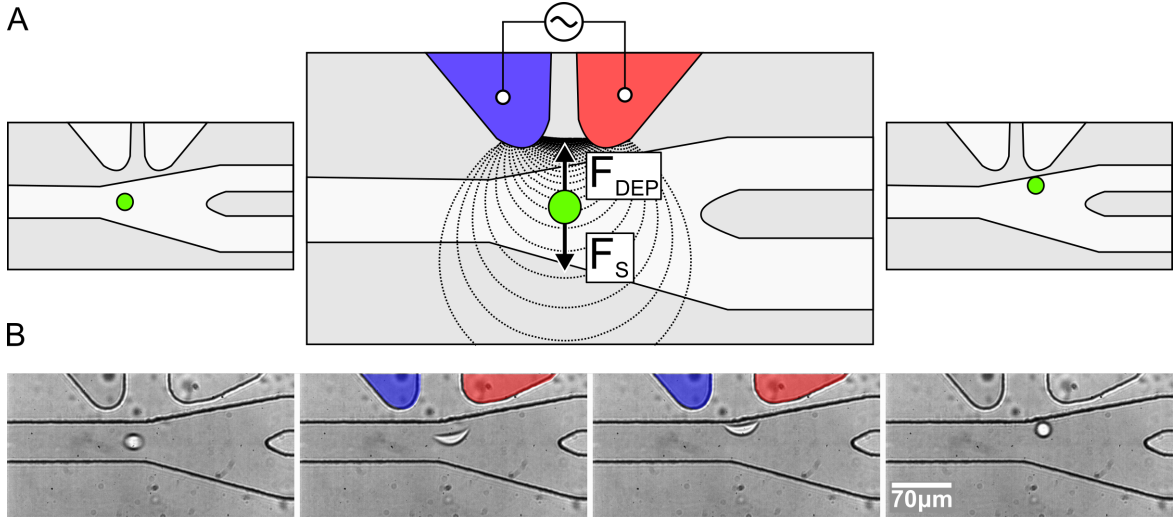


Fig. 2.1 Dielectrophoresis is used to sort droplets. A: The schematic shows how a droplet is laterally displaced by switching on the electrodes (indicated by blue-red filling). In positive dielectrophoresis the force \vec{F}_{DEP} points along the gradient of the electric field \vec{E} (dashed lines). The dielectrophoretic force is balanced by the Stokes drag \vec{F}_S . B: Snapshots of a video showing how a droplet is deviated from its trajectory by switching on the electrodes. Note that a high voltage V was applied to demonstrate the vertical displacement. For routine sorting, lower voltages that do not deform the droplets are used.

dielectrophoretic force is balanced by the Stokes drag, given by:

$$\vec{F}_S = 6\pi\eta_{\text{oil}}r\vec{v} \quad (2.3)$$

With η_{oil} being the viscosity of the oil and \vec{v} the speed of the droplet. Thus, and with the vector transformation $2(\vec{E}_{\text{rms}} \cdot \nabla)\vec{E}_{\text{rms}} = \nabla \vec{E}_{\text{rms}}^2$, [106]:

$$\vec{v} = \frac{\epsilon_{\text{oil}}r^2\nabla \vec{E}_{\text{rms}}^2}{3\eta_{\text{oil}}} \quad (2.4)$$

Which can be simplified to [103]:

$$\|\vec{v}\| = \frac{\epsilon_{\text{oil}}r^2kU^2}{3\eta_{\text{oil}}} \quad (2.5)$$

Where k is a geometric factor describing the electrode shape and location and U the applied voltage. Therefore, smaller droplets move at lower speed towards the electrode, making it more difficult to collect them into the sorting channel. Also, a carrier oil with low viscosity and high permittivity is beneficial to sorting efficiency, which is one advantage of the now commonly used perfluorinated oils over hexadecane ($\epsilon = 6$ and $\eta = 1$ mPa s for HFE7500; $\epsilon = 2$ and $\eta = 8$ mPa s for hexadecane [103, 107]).

Combining dielectrophoretic actuation with a fluorescence detection system, yielded the first fluorescence-activated droplet sorter [60]. Baret *et al.* demonstrated the enrichment of cells harbouring a functional β -galactosidase gene over cells harbouring an inactive one. The enrichment was performed at 300 Hz, where the sorter had a minimal error-rate of 10^{-4} . Droplet sorting was possible up to 2 kHz albeit at higher error-rates. This study was soon followed up by a directed evolution experiment [64]. The sorting chip was improved, run at 2 kHz, and 10^7 cells were sorted on a single day. This showed the utility of droplet microfluidics to access a much larger proportion of a library's diversity. The record sorting rate for FADS is now 30 kHz matching the rates of commercial FACS systems [108]. The method has also been extended to sort droplets based on absorbance measurements, albeit at only 300 Hz [61].

In the following section, I will describe the FADS which I built following the principles laid out here and starting from the principal set-ups published in [60, 67]. A FADS can be separated into two main components: the instrumentation to run the sorting chip and the sorting chip itself. The goal my work was to establish the instrumentation rather than optimising the chip design. The primary chip design used was previously published in [67].

2.3 Description of the FADS set-up

The FADS system I developed measured green fluorescence induced at 488 nm, analysed the signal, sorted droplets, and monitored the sorting performance using high-speed videos. What follows is a description of what I achieved as of September 2018. First the optical and electronic set-up and then the sorting algorithm will be explained.

2.3.1 The optical and electronic set-up

The optical set-up was optimised for the detection of fluorescein (excitation maximum at 492 nm and emission maximum at 512 nm [109]), which was used as the reporter molecule of choice in several studies performed using this instrument.

Fluorescence was induced at 488 nm and measured at 525/28 nm, while the performance of sorting was monitored using red light above 593 nm. The arrangement is depicted in Figure 2.2A¹. The microfluidic sorting chip was illuminated by a white halogen lamp (100 W, BF) *via* a longpass filter F1 (F1, 593 nm) to prevent green illumination light from reaching the detector. The laser beam (488 nm, 30 mW, attenuated to 3 mW with 1.0 ND filter) was focused onto the chip through an objective (LUCPlanFLN 40x/0.6, Olympus). The induced fluorescence and red light were directed past a longpass filter (F2, 488 nm) to a second dichroic mirror DC2 (550 nm). The red light was captured by a high-speed camera (CAM). Lower wavelengths were directed to a band pass filter (F3, 525/28 nm) before reaching the photomultiplier tube (PMT) for detection. Typically, the gain of the PMT was set to 0.7 V (maximum of 1.25 V). The PMT output voltage was read by an field-programmable gate array (FPGA), which I programmed to measure the fluorescence signal of each droplet. If a droplet fulfilled the operator-set sorting criteria, the FPGA triggered the pulse generator (PG) and video acquisition by the high-speed camera (CAM).

The sequence of voltage pulses for sorting is shown in Figure 2.2B. The FPGA sent a 50 μ s transistor-transistor logic (TTL)² trigger to the pulse generator which opened a gate for the function generator (FG). The function generator then produced a square-wave which was amplified a hundred times (AMP) and applied to the electrodes embedded in the microfluidic chip. The width of the gate could be set by the operator. The standard width was 500 μ s for 2 pL droplets and 5 ms for 65 to 200 pL droplets. The frequency of the square-wave was set to 10 kHz with an amplitude between 6 and 9 V depending on droplet behaviour. This cascade

¹The spectra of the filters are given in Figure A.1 and the model specifications in Section 7.1.4.

²In TTL there are two states: *low* and *high*. A trigger sets the state to *high* for a brief period. Standard TTL circuits operate at 5 V. A TTL input signal is defined as *low* between 0 and 0.8 V and *high* between 2 and 5 V.

Table 2.1 Standard droplet sorting parameters for 2 pL droplets. Acronyms refer to Figure 2.2.

Paramter	Value
Oil flow rate	100 - 150 μLh^{-1}
Droplet flow rate	5 - 15 μLh^{-1}
CAM frame rate	4 kHz
CAM exposure time	20 μs
FPGA sampling rate	200 kHz
FPGA trigger width	50 μs
FPGA trigger voltage	5 V
PMT Gain	0.7 V
PG Voltage	5 V
PG Period	1.0 ms
PG Width	0.5 ms
FG Voltage	6 V
FG Frequency	10 kHz

of pulses resulted in a selected droplet being dielectrophoretically pulled into the collection channel.

The process was continually monitored by the high-speed camera which was typically acquiring at 4,000 frames per second. When the camera was triggered by the FPGA, a video file was saved from *e.g.* 10 frames before to 20 frames after the trigger signal. As shown in Figure 2.2 this enabled the manual inspection of every single sorting event and the adjustment of the sorting parameters to optimise sorting. Video acquisition was also used as a tool to check for inadvertently collected droplets due to errors, which could lead to false positives in screening campaigns. In general, droplet sorting worked *ad hoc* if the indicated parameters were used (summarised in Table 2.1). However, differences in droplet size and content between experiments sometimes required small adjustments at the start of a run.

2.3.2 The sorting algorithm

In a previous iteration, the fluorescence signal was split into two lines. One was wired to the input of a microprocessor which made the sorting decisions, the second was used to record, analyse, and visualise the signal using a custom LabView virtual instrument (VI), which I designed³. The splitting had been necessary because the microcontroller could not stream the analysed data to the host (a desktop PC) and make sorting decisions simultaneously. This limited the ability to precisely monitor and control the actions of the microcontroller.

³LabView is a systems engineering software used to control hardware by National Instruments.

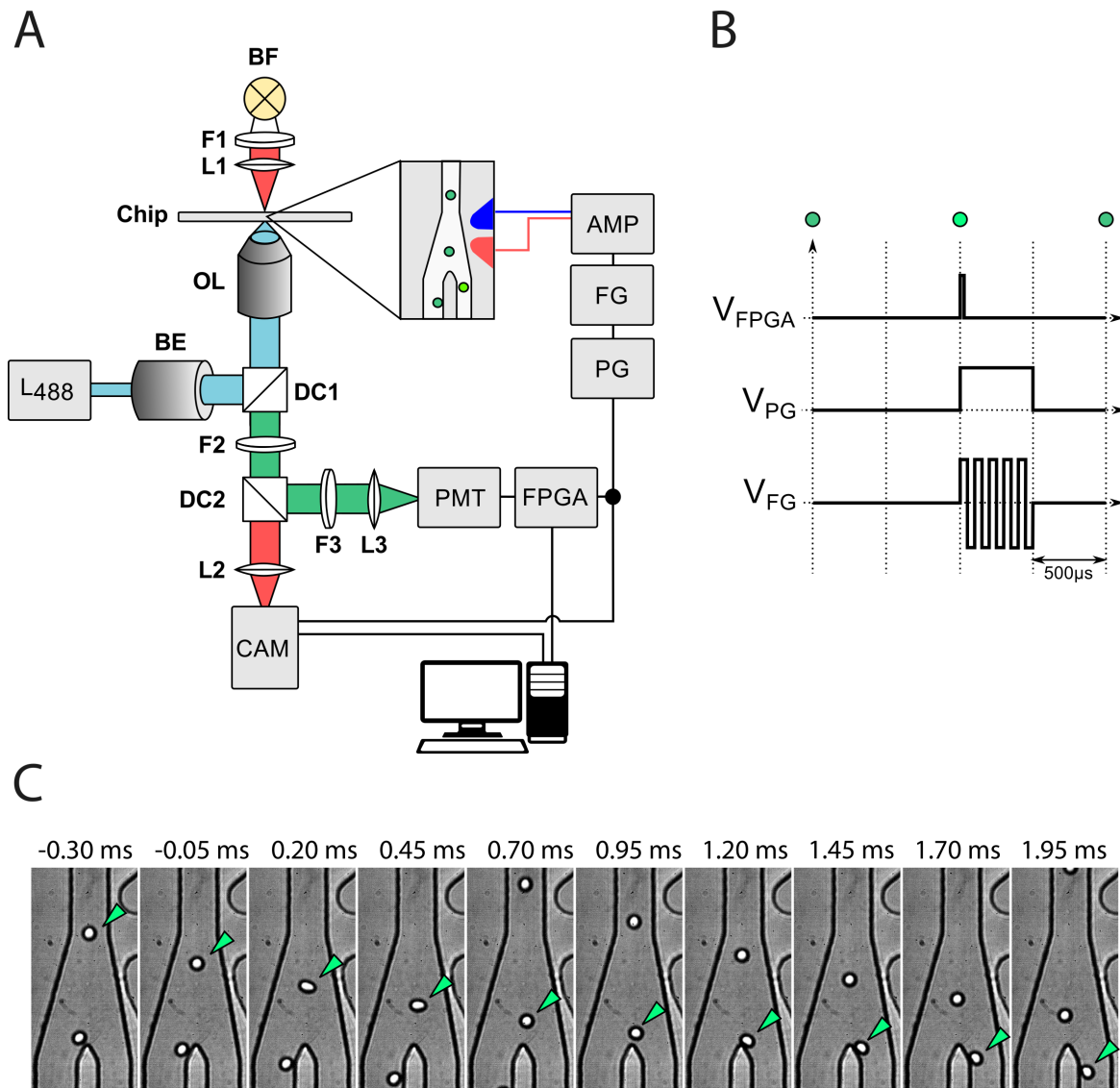


Fig. 2.2 The figure shows how the FADS achieves droplet sorting. A: The optical set-up. BF: bright field, F: filter, OL: objective lens, BE: beam expander, DC: dichroic mirror, CAM: camera, PMT: photomultiplier tube, FPGA: field-programmable gate array, PG: pulse generator, FG: function generator, AMP: amplifier. The inset shows how the chip is connected to the amplifier. The spectra of the filters are shown in Figure A.1. B: The pulse sequence was started by the FPGA, which results in the sorting of a droplet. C: Individual frames of a video recorded by the camera for one sorting event. Droplets entered from the top and exited *via* the waste channel on the left. The droplet which triggered the sorting event was pulled to the collection channel on the right as indicated by the green arrows.

In the latest implementation of the FADS, the fluorescence signal was processed by an FPGA. I programmed the FPGA using the LabView FPGA module allowing the smooth interfacing between the FPGA VI and the host VI such that control, execution, and monitoring of the sorting process were synchronised. This was due to the chief advantage of an FPGA: it can execute different processes in parallel. This also allowed the calculation of several droplet properties at high speed enabling droplet sorting based on more than one criterion.

The four droplet properties that were calculated were the peak amplitude U_{max} , the area under the peak, the peak WIDTH at the threshold, and the full width at half maximum (FWHM). The width measurements (in terms of time spent within the laser beam) were introduced to distinguish droplets of the correct size from satellites and larger droplets, which was necessary for some experiments performed on the set-up (see Section 2.4.2). The FWHM was introduced as a better measure for droplet size compared to the WIDTH. It was assumed that each droplet's signal could be approximated using a normal distribution:

$$f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-t_c)^2}{2\sigma^2}\right) \quad (2.6)$$

Where σ is the standard deviation, t is time, and t_c the center of the distribution. Under this assumption, the FWHM depends linearly on σ , which was practical to sort droplets based on their size. The WIDTH at the threshold was dependent both on σ and on U_{max} limiting its use for size selections as will be shown in Section 2.4.2.

Figures 2.3 and 2.4 show how the four properties were calculated. The FPGA VI was executed every 5 μ s resulting in a sampling rate of 200 kHz. At each iteration of the program, the sorting parameters and one data point $U(t)$ were read and sent to downstream processes. Each unprocessed data point was also added to a FIFO buffer⁴ which was read by the host VI. At 200 kHz this yielded, for example, 40 datapoints for a droplet that spent 200 μ s in the laser beam. In Figure 2.3 the four critical points at which the FPGA started or stopped a process are indicated with red circles. A droplet was assumed to be passing through the laser, if the signal was above an operator-set threshold. When threshold was crossed upwards (rising edge) the program started calculating the droplet properties. As the rising edge was passed, the time t_1 was stored and two processes started. The area under the signal was approximated according to:

$$A(t) = A(t - \Delta t) + U(t) * \Delta t \quad (2.7)$$

⁴A FIFO (first in, first out) is a data buffer where the first element added to the buffer is the first element read and removed. Refer also to Figure 2.5.

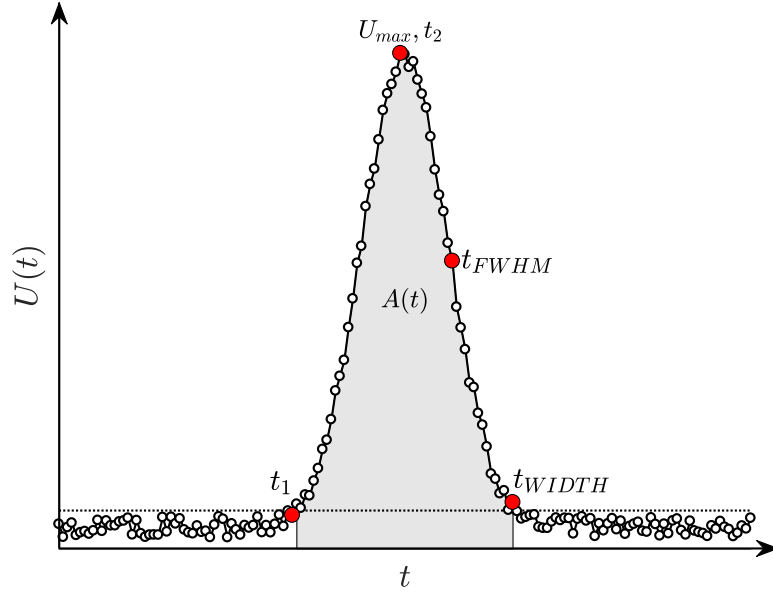


Fig. 2.3 Simulated fluorescence signal using a normal distribution with added normally-distributed noise. White circles indicate each datapoint $U(t)$ read by the FPGA. Red circles indicate moments at which the FPGA starts or stops a process. Dashed line: threshold, t_1 : time of rising edge, U_{max} : maximum signal at t_2 , t_{FWHM} : time of half maximum, t_{WIDTH} : time of falling edge, and $A(t)$: area under the signal.

Where $A(t)$ is the area at the current iteration, $A(t - \Delta t)$ the area from the previous iteration, and Δt the time between iterations (fixed to $5 \mu s$). The second process checked if the incoming datapoint $U(t)$ was smaller than the previous one $U(t - \Delta t)$:

$$U(t) < U(t - \Delta t) \quad (2.8)$$

If this was true, the previous datapoint was stored as the amplitude U_{max} . The process remained active until the falling edge was detected, *i.e.* it determined the global maximum between the two edges. A state change of Equation 2.8 from false to true started a third process. This process determined the FWHM by storing the time t_2 at which the maximum was reached and starting the comparison:

$$U(t) \leq \frac{U_{max}}{2} \quad (2.9)$$

When this comparison became true, the difference $t_2 - t$ yielded the half width at half maximum, which was multiplied by two to obtain the FWHM. Note that this process would reset each time U_{max} was updated, *i.e.* until the global maximum was reached.

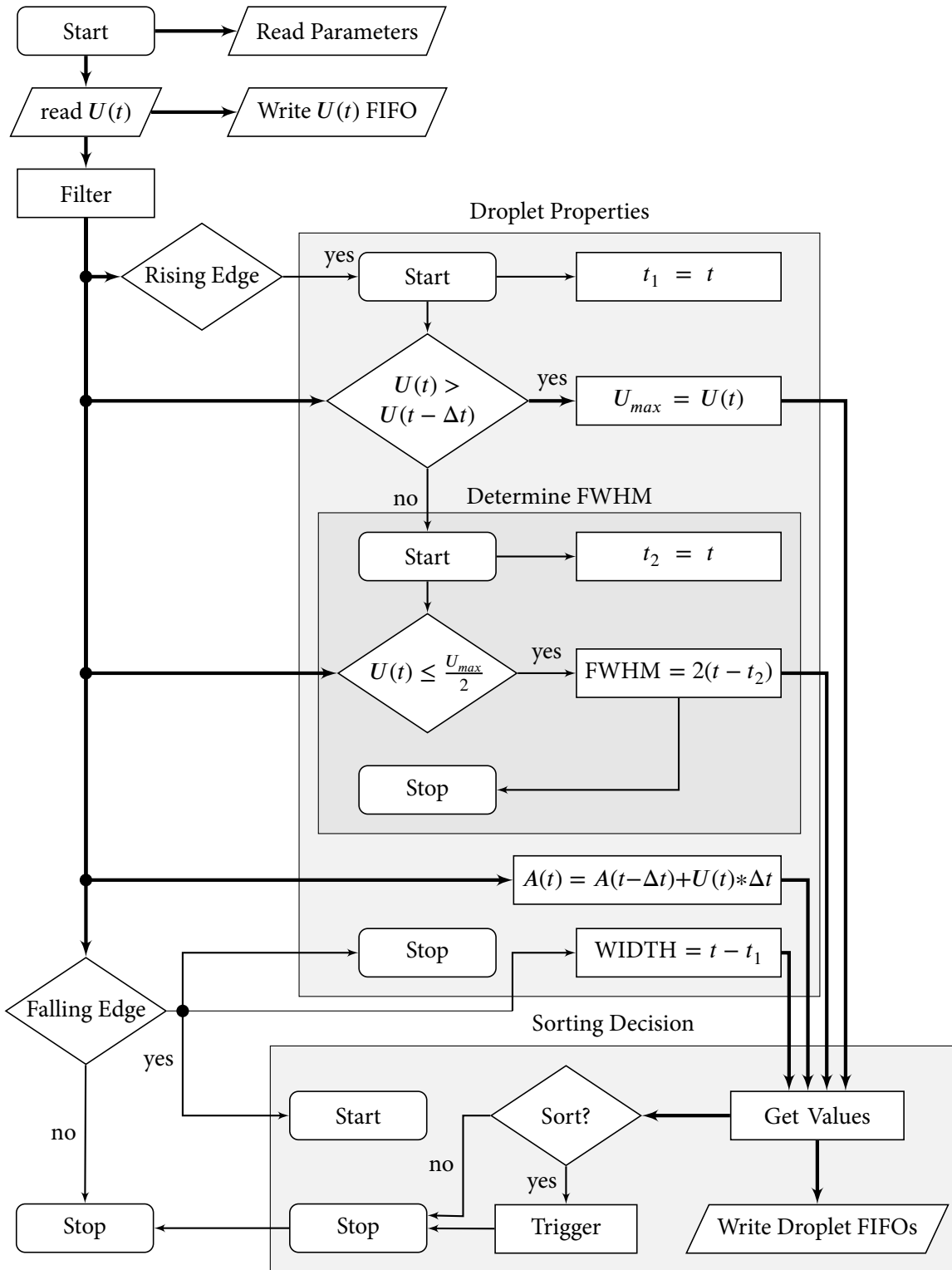


Fig. 2.4 Flowchart of the signal processing implemented on the FPGA. Tasks branching off at nodes are executed in parallel. The processes in the droplet properties block are started when a rising edge occurs and stopped when the falling edge follows. The calculated values are then used to make a sorting decision. Thick lines indicate the passing of data, thin lines instructions to execute a step.

Finally, as the falling edge was passed, the difference $t_1 - t$ yielded the WIDTH at the threshold level. At this point, all analysing processes were stopped, and the obtained values compared against the sorting parameters, the sorting decision was made, and the values added to one of two FIFOs, one for the sorted and one for the unsorted droplets.

The data transfer between FPGA and the host was handled by the direct memory access (DMA) engine provided by LabView as shown in Figure 2.5. For each FIFO on the FPGA side there was a corresponding one on the host side. Data points were transferred one-by-one from the FPGA to the host buffer in a synchronised fashion to ensure there was no data loss.

Before the host VI read the data from its buffers, it updated the sorting parameters as shown in Figure 2.6. This was achieved by directly updating the hardware registers holding each value on the FPGA device. The sorting parameters that an operator could adjust were the threshold for peak detection, the maximum and minimum peak amplitude, and a choice between either the maximum and minimum FWHM or WIDTH. The program then read, analysed, and plotted the data from different FIFOs. Three key indicators were the rate of droplet sorting, the total number of droplets analysed, and the number of droplets collected. Finally, the signal properties for each sorted droplet were logged in a separate file, before the program was executed again.

This analysis workflow improved droplet sorting, because the droplet size criteria allowed the exclusion of droplets of incorrect size (due to merging or splitting). The background signal can be different in droplet of differing size, therefore both split and merged droplets can appear as shoulders in the fluorescence (amplitude) histogram and overlap with the signal of real hits. If the frequency of hits in the droplet population is smaller than that of such events, *e.g.* 10^{-5} and 10^{-4} , this can lead to the collection of a large amount of false positives. Therefore, assays can significantly profit from information on droplet size.

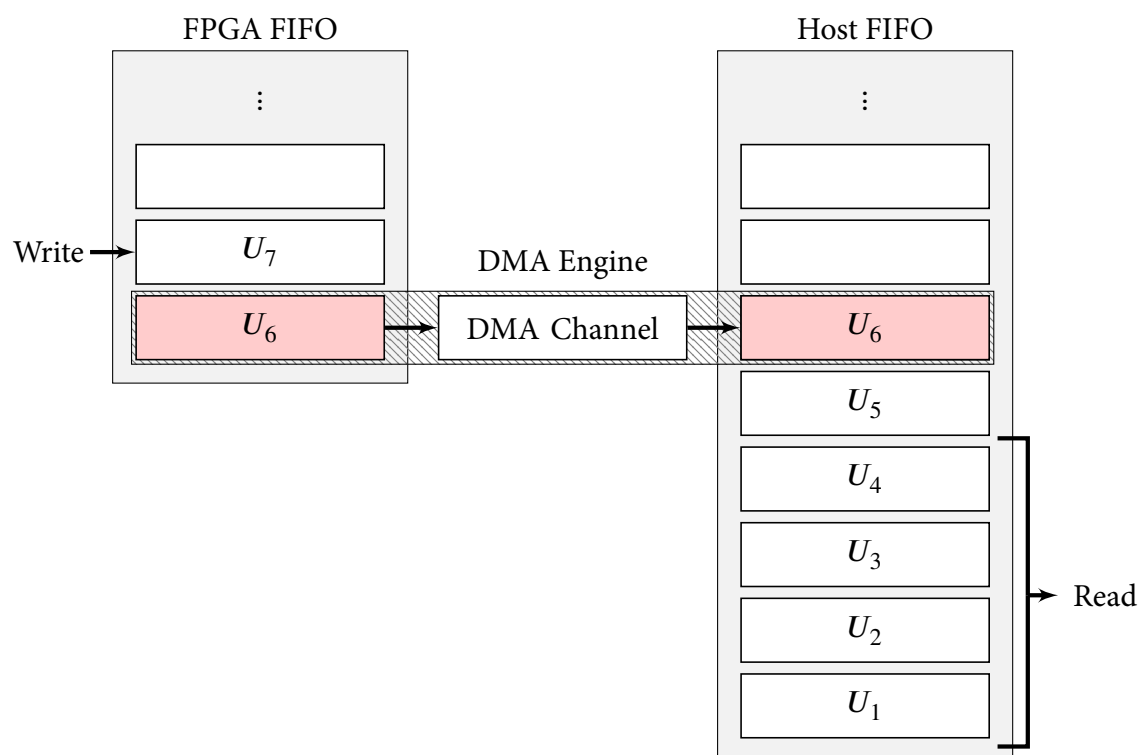


Fig. 2.5 The DMA Engine transferred data between the buffers on the FPGA and on the host. While the FPGA wrote one element at a time at a high rate (200 kHz for the raw signal, at most 8 kHz for droplet data), the host read data more slowly (10 Hz) but in larger batches, thus ensuring no data were lost.

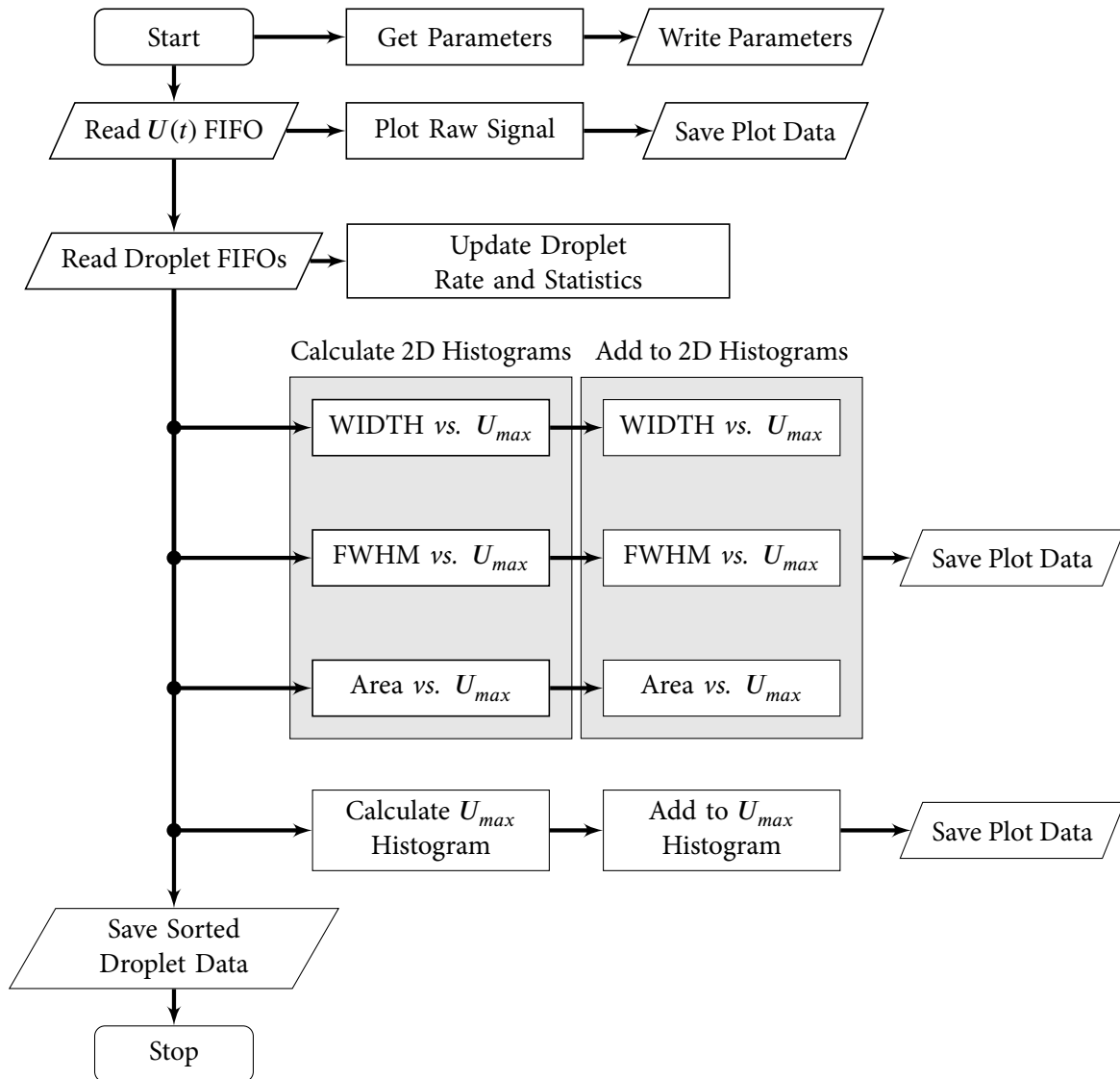


Fig. 2.6 Flowchart of the LabView program used to monitor droplet sorting. The data were read from the FIFO buffers and plotted. The tasks branching off at nodes were executed individually before the program moved to the next branch point. Sorting criteria could be defined according to the obtained data and changed at any point. The sorted droplet data were logged separately.

2.4 Results and discussion

Experimental results obtained from the FADS are reported in the next chapter. Here, the aim is to illustrate how the set-up was tested and sequentially improved to reach its final state described above.

2.4.1 Single-photon counting module *versus* photomultiplier tube

Two types of detectors were compared, a single-photon counting module (SPCM) and a PMT. The former is used for ultra-high sensitivity single-molecules studies and the latter is widely used in plate readers and FACS instruments. While PMTs have previously been used in FADS, I tested the SPCM to assess whether it was suitable for sensitive fluorophore detection in droplet sorting applications.

The SPCM required adjustment of the signal processing algorithm. While the output of the PMT was an analogue voltage signal between 0 and 10 V, the output of the SPCM was digital. It sent one TTL pulse per detected photon to the FPGA. The detection efficiency in this wavelength range was about 50% and the maximum photon counting rate was 39 MHz. A counter was implemented on the FPGA which monitored the SPCM at 40 MHz, which was then sampled analogously to the voltage signal. The sampling rate for droplet detection and analysis was set to 50 kHz during this experiment. Thus, the signal range of the SPCM was from 0 to 800 photons, an overview is given in Table 2.2. Peak analysis was performed in the same way for both detectors.

Table 2.2 Comparison of some parameters of the SPCM and PMT as implemented here.

Parameter	SPCM	PMT
Flexible Gain	No	Yes
Sensor diameter	0.18 mm	22 mm
Quantum efficiency [†]	70%	21%
Band width	N/A	0 - 20 kHz
Maximum counting rate	39 MHz	N/A
Output format	digital	analogue
Output range	0-800 photons	0 - 10 V
Output linear	< 40 photons	≤ 10 V

[†] at respective peak wavelength.

N/A - not applicable.

A single-cell esterase assay in 15 μm droplets was performed using a positive control to test the detectors under real conditions⁵. Not all droplets were expected to contain cells due to the Poisson distribution governing cell encapsulation, *i.e.* two peaks were expected: one with low background fluorescence (65% of droplets) and one with increased fluorescence due to enzymatic activity (35% of droplets). The sample was injected two hours after generating the droplets. At this incubation time, the separation between positive and negative droplets was not expected to be large. The two measurements were performed one after the other by simply switching the detector.

The results are presented in Figure 2.7. As can be seen in the width *versus* amplitude plots, the population of droplets was more polydisperse at the beginning of the measurement (top panel), compared to the end (bottom panel). This was typically observed in droplet experiments, most likely because droplets flowing at the front of the tubing used to connect devices experienced pressure fluctuations when connecting and disconnecting to microfluidic devices. When selecting droplets in the expected size range (120 to 320 μs) some outliers are removed, but the shape of the overall amplitude histogram does not change significantly.

Importantly, both detectors were able to discriminate the two expected peaks. The difference in signal between the peaks was 2 fold for both detectors. The main distinction between the two detectors is in the right-hand side tail of the distributions: the SPCM histogram ends abruptly while the PMT histogram tails off. The latter behaviour is typically observed in single-cell experiments and can be explained by the varying amount of protein expressed by individual cells.

The difference observed for the SPCM can be explained by a deviation from linearity at a high rate of incident photons. Above a counting rate of 1 to 2 MHz (photon count of 30 to 40), the detector starts to underestimate the real rate. At a readout of 17 MHz (photon count 350), the actual rate of incident photons is two-fold higher. Thus, under these conditions the amount of light emitted by the droplets was too high to ensure the linearity of the measurements. At a higher attenuation of the laser, the signal was too low to observe the empty droplets.

⁵The esterase used here was Est30, which was used to establish the esterase assay in droplets. Refer also to Section 3.3, specifically Table 3.1. Reaction conditions as in Section 7.1.6

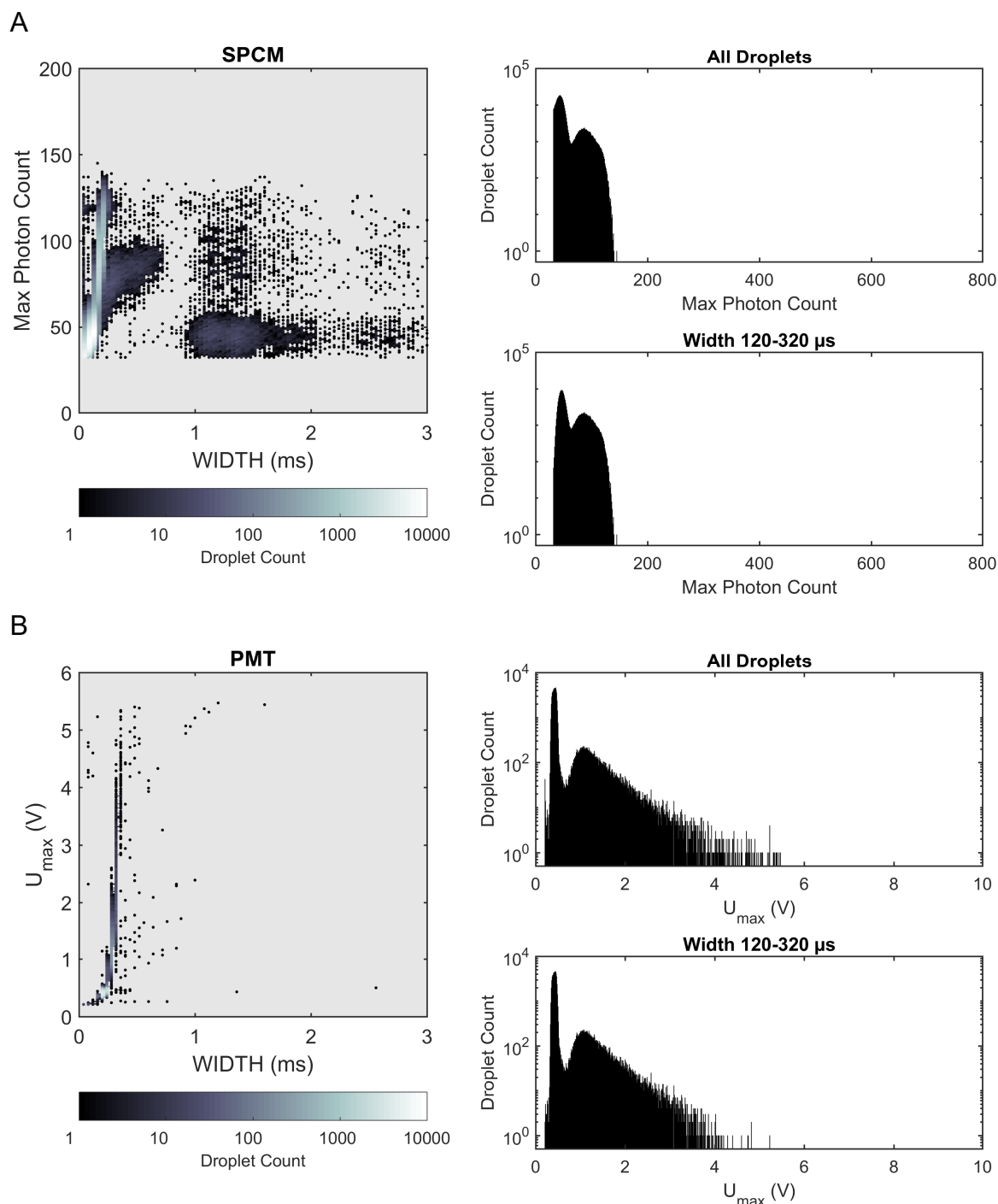


Fig. 2.7 The two panels show the difference in using two different detectors to measure the fluorescence distribution of the same droplet sample. *A*: single-photon counting module (SPCM). *B*: photomultiplier tube (PMT). As opposed to the PMT, the dynamic range of the SPCM did not allow an accurate measurement of droplets with higher fluorescence.

It was therefore concluded that the PMT was better-suited for this application. This was practical for two further reasons. First, the sensor diameter of the PMT is larger than for the SPCM (22 mm vs. 180 μm), simplifying the alignment and maintenance of the optical set-up. Second, while the SPCM had a fixed gain, the gain of the PMT could be adjusted to fine-tune the sensitivity as required by each type of experiment.

While the PMT was preferable to the SPCM, there was still a trade-off between sensitivity and sorting speed due to its limited bandwidth. In this case, it was 0-20 kHz, *i.e.* a time-resolution of 50 μs . At a sorting rate of 2 kHz, the centre of droplets may pass the laser every 500 μs , but because of the required spacing between them, each one may only spend 100 μs within the illumination area, which means at higher sorting rates the measured fluorescence signal is at risk of becoming attenuated. In future, new detectors with both high speed and high sensitivity are needed to improve fluorescence detection in droplet sorters.

Next, the linearity and sensitivity of the PMT was measured under realistic sorting conditions, *i.e.* at droplet rates of several hundred hertz and using 55% of the maximal gain voltage. Droplets were generated containing fluorescein at 8 nM to 1 μM in 50 mM TrisHCl pH 8 and measured, see Figure 2.8. Using linear regression the concentration at the detection threshold in this experiment was (13 ± 3) nM. Given the droplet volume of 1.8 pL, a concentration 16 nM corresponds to about 10^4 fluorescein molecules. This would require only 100 enzyme molecules to perform 100 reaction turnovers. Even weakly expressed, promiscuous enzymes are expected to surpass these numbers. Therefore, the system should enable highly sensitive screening of metagenomic libraries.

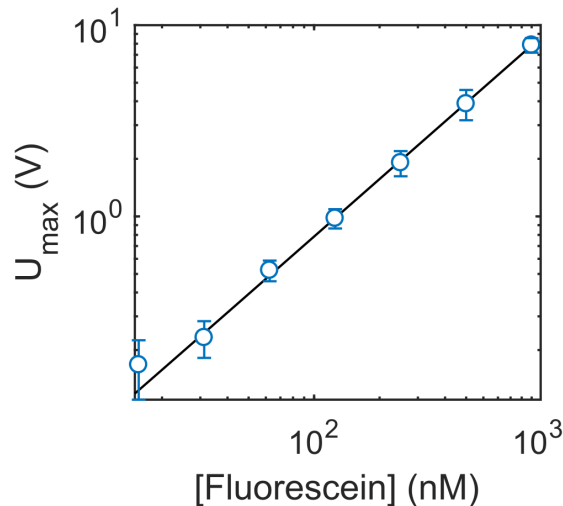


Fig. 2.8 Linearity of the fluorescence measurement using the PMT and fluorescein in 15 μm droplets as a standard. Error bars correspond to the FWHM of the droplet population. The solid line shows a linear regression ($R^2 = 1$).

2.4.2 Improvements to the FADS were driven by research projects

The FADS was successfully used in a number of projects. These projects that have yielded improved or new enzymes to date are listed in Table 2.3. I performed the droplet sorting in the protease project, established the droplet reaction conditions and performed the sorting in the sulfatase autodisplay project, and supported all other projects. The metagenomic screening for esterases was a major part of this thesis and is reported in detail in Chapter 3. Here, I will briefly highlight how the FADS contributed to the success of the first two projects listed in the table.

Dr Josephin Holstein performed a directed evolution campaign of a protease in droplets. Proteases are difficult to evolve in *E. coli*, because they are toxic to the host cell. Other hosts, such as *B. subtilis*, are more tolerant, but suffer from low transformation efficiencies limiting library size. Therefore, these enzymes are a case where cell-free expression systems are of high interest. Christian Gylstorff established a cell-free expression system in droplets based on *in vitro* transcription and translation (IVTT) [110]. Building on his work, Dr Josephin Holstein then performed a directed evolution campaign. However, the implemented workflow led to the generation of unwanted satellite droplets that did not contain functional genes and limited the sorting efficiency for this assay. Because the formation of these satellites could not be prevented, the sorting of droplets based on both peak amplitude and width was developed.

Figure 2.9 shows how the width measurement helped to improve the outcomes of this project. As can be seen in the left panel, the expected droplet size was about 1.1 ms, while the satellites were at 120 to 240 μ s. Therefore, only droplets with a width between 0.65 and 1.5 ms were selected for sorting. As the two histograms show, this reduced the number of events at

Table 2.3 Overview of the projects successfully performed on the FADS.

Enzyme	Expression	Droplet Diameter (μ m)	Collaborator
Directed Evolution			
Protease	IVTT [†]	50	Dr Josephin Holstein
Sulfatase	<i>E. coli</i> Autodisplay	15	Dr Bert van Loo
Phosphotriesterase	Single-cell lysate	20	David Fernández Schnettler
Metagenomics			
Esterase	Single-cell lysate	15	Dr Mariana Rangel Pereira
Glycosidase	Single-cell lysate	20	Stefanie Neun
Sulfatase	Single-cell lysate	20	Stefanie Neun

[†] *in vitro* transcription and translation

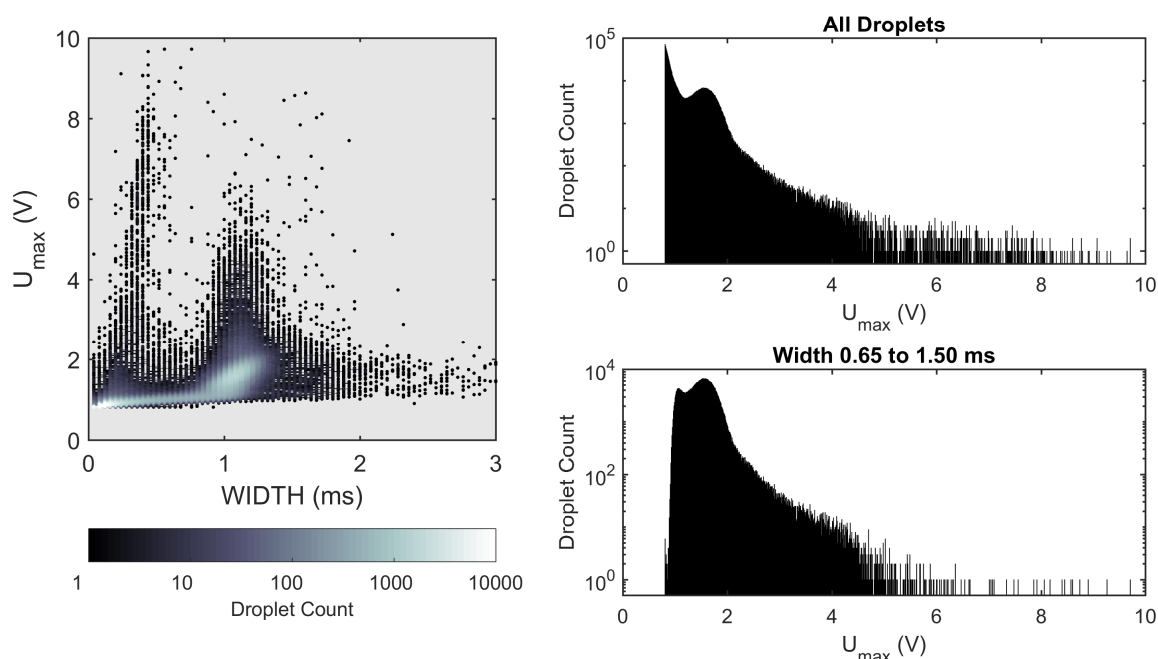


Fig. 2.9 Screening of a protease library screen using a cell-free expression system. The workflow creates small droplets, which could be separated visually from the main droplet population using the width measurement. Sorting based on both width and amplitude of the droplets improved the outcome of screening campaigns like this.

amplitudes above 5 V. This is important, because often only a certain number of droplets can be collected due to the limited throughput of the available methods for downstream characterisation of the selected variants. Before the width measurement was implemented, the sorting threshold was set too high and only few genuinely improved variants would have been selected. Furthermore, differing droplet sizes prevent regular spacing due to a difference in flow resistance [111]. Therefore, small droplets caught up with larger droplets and caused them to be sorted as a false positive. Avoiding any small droplets from triggering a sorting event therefore reduced the false positive rate. Using this improved sorting, it was possible to sort six separate libraries and obtain 38 improved enzymes compared to the starting point.

In another project, I supported Dr Bert van Loo in evolving a sulfatase in droplets using an expression system in which the enzyme was presented to the extracellular space by *E. coli* [112]. This system avoided cell lysis, which allowed live cell recovery and thus simplified the confirmation and characterisation of selected variants.

During the project it became evident that the first implementation of the width measurement was not independent of the droplet amplitude. Figure 2.10 shows a control experiment illustrating this point. A positive control and a negative control were encapsulated. Although almost all droplets containing the positive control were saturating the detector, there was a

spread of signal amplitudes across the width-range. Even though the droplets were uniformly sized according to video analysis, the width *versus* amplitude plot implied a more than three-fold difference in size between the least and most fluorescent droplets.

A simple simulation of the width measurement based on the normal distribution showed that this was indeed an artefact of measuring the width at a fixed threshold (see Appendix Figure A.3). Therefore, the determination of the FWHM was implemented, which measured the width relative to the amplitude of the droplet. Thus, an amplitude-independent measure of droplet size was created, see Figure 2.10B.

Taken together, these two examples illustrate how the performance of the FADS was incrementally improved based on experimental needs.

2.5 Conclusions

The set-up which I have developed in this chapter is a state-of-the-art instrument to control droplet sorting using a microfluidic chip. It has a proven ability to sort droplets for a number of biochemical procedures using several droplet properties.

In its current state, the FADS can only measure green fluorescence. Yet, it provides a tested framework for extension to additional fluorescence measurements by incorporation of more lasers and PMTs. Work by Dr Tomasz Kaminski is currently on the way to achieve this. The peak analysis algorithm can be used as a module to analyse each fluorescence channel separately. Beyond access to different fluorescent dyes, plotting the output from two fluorescence channels can be used to control for expression variation between cells by co-expressing, for example, a red fluorescent protein from a plasmid. Also, a strain that constitutively expressed a red fluorescent protein from its genome could be used to control for the number of cells if live cells are used instead of lysed cells. Such measurements would improve the outcome of droplet sorts.

There are improvements that can be made to the peak analysis algorithm. The current way to determine the amplitude U_{\max} is the simplest. However, the time at which the global maximum occurs may be different from the true centre of the peak ($t_2 - t_c \neq 0$). This can be caused by inhomogeneous droplet contents due to, for example, a co-encapsulated dust particle. In such a case, both the width and the FWHM measurement are inaccurate. To resolve this, all data-points could be stored in an array and, once the falling edge occurs, several values around the temporal midpoint could be averaged to obtain a more robust measure of the amplitude. Similarly, the FWHM could be determined as the difference between the temporal midpoint and the time point at which Equation 2.9 becomes true.

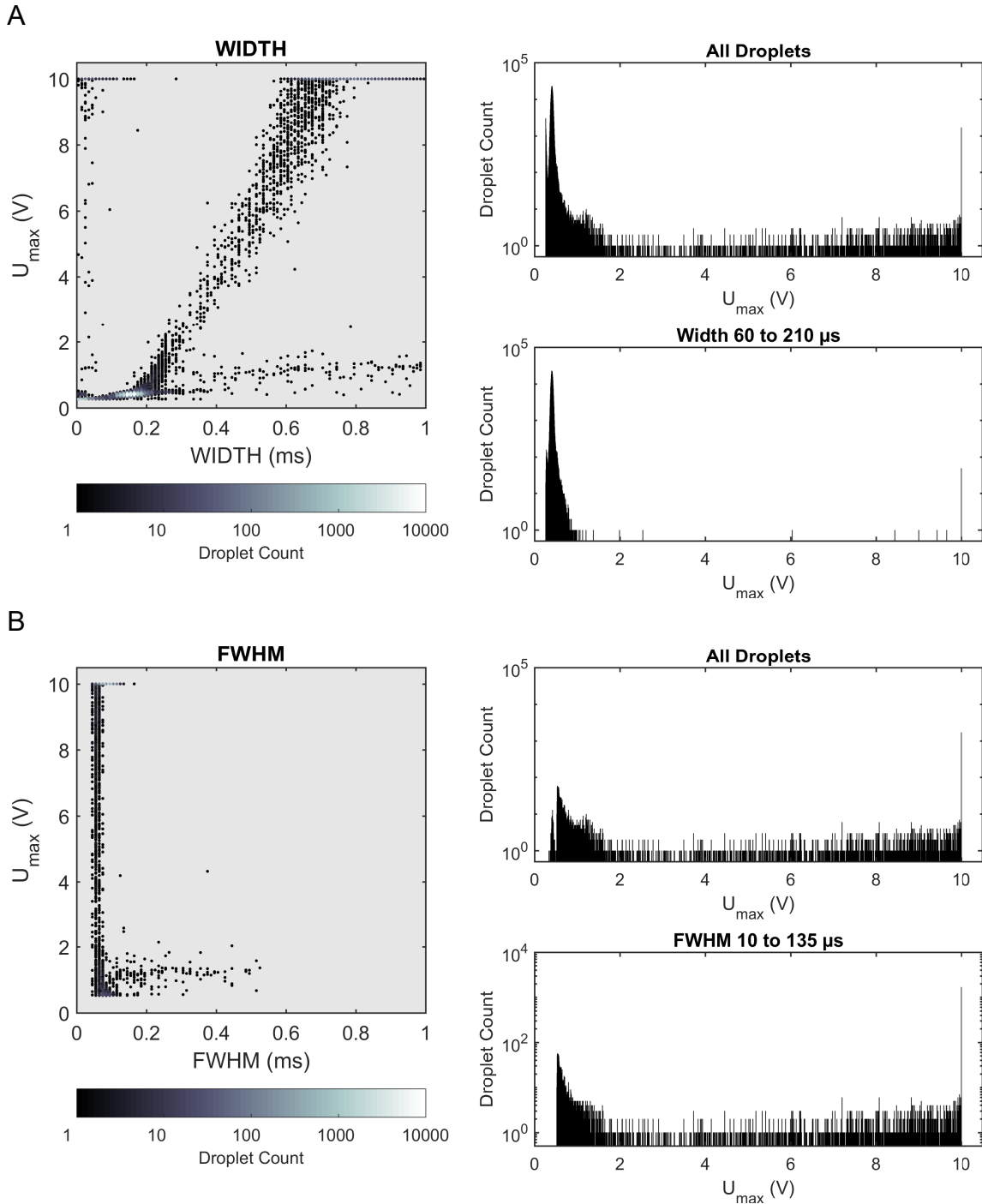


Fig. 2.10 The difference between the WIDTH (A) and the FWHM (B) measurement of droplet size for the same sample. The WIDTH measurement is dependent on the amplitude of the signal because it is measured where the peak passes the detection threshold, whereas the FWHM measures the width at half the amplitude of a fluorescence signal. Therefore, gating according to droplet size is more reliable using FWHM.

Another, more difficult to resolve, limitation of the algorithm is that if FWHM is below the threshold level, it cannot be determined. Therefore, it may be necessary to offset the fluorescence by adding a small amount of fluorescein (*e.g.* 10 nM) to the buffer.

The sensitivity of the previously reported set-ups was 1 nM for amplex red and 2.5 nM fluorescein with droplets of 23 and 20 μm respectively [64, 67]. These were determined using the same concentration of fluorophore in all droplets. In this chapter, smaller, 15 μm , droplets were used to reduce the dilution of enzyme upon cell lysis. In a mixture of droplets with different fluorophore concentrations, the lowest concentration which could be distinguished from background at medium gain was 16 nM fluorescein. Because the fluorescence measurements are performed with droplets under flow, smaller droplets will emit fewer photons in total, because they spend less time in the laser beam, explaining the somewhat reduced sensitivity.

The sorting and error rates were not discussed in this chapter, because they are mostly determined by the chip design [108]. False positives in general may be dominated by biological factors. This will form part of the discussion in the next chapter. In it, I will present the application of the discussed set-up in a functional metagenomic screen for esterases.

Chapter 3

Functional metagenomic screening for esterases in droplets

3.1 Abstract

In this chapter I present a functional metagenomic screen for esterases in microfluidic droplets. Esterases and lipases are an important class of enzymes used in industry . They have been extensively screened for in metagenomic libraries, albeit using cost- and time-intensive culture plate methods with most studies reporting only one or two new enzymes. Interestingly, no major study published to date has used droplet sorting to screen for esterase activity either to improve existing or discover new esterases. Yet, such a screen is highly desirable to increase throughput and reduce the cost of screening.

Here, I met this challenge by developing an esterase assay in droplets and using the droplet sorter I built in Chapter 2 to screen the million-member metagenomic library SCV. The throughput of droplets allowed full coverage of the library on a single day of screening. Therefore, this is the largest functional metagenomic esterase screen performed to date in terms of throughput. Thirteen new genes associated with esterase activity were identified. Most of these belonged to the α/β -hydrolase superfamily of proteins, which contains most of the known esterases. However, many were rare members of small families such as N20 and RR11ORF1, which are members of the recently discovered family DUF3089. The enzymes were purified and quantitatively characterised in terms of their esterase activity. They were also tested against a range of other substrates to find promiscuous activities. The hits N1O5 and N7 showed β -galactosidase and Kemp eliminase activity, respectively. This shows that using one substrate as “bait” can elicit enzymes that are sometimes difficult to screen for be-

The results presented in this chapter were obtained in collaboration with Dr Mariana Rangel Pereira.

cause of lack of sensitivity or high background. Together, this research enabled the discovery of new esterases from small protein families at high throughput and low cost.

Contributions: *I performed all of the microfluidic work: I tested and optimised the reaction conditions in droplets, I expressed and prepared the metagenomic library for droplet screening, screened the droplets, recovered and re-transformed the DNA. I performed the re-screening and identification of hits on tributyrin plates together with Dr Mariana Rangel Pereira. The sequence analysis, family assignment, sequence-similarity network, and interpretation of findings presented here is entirely my work. The protein expression and determination of the esterase kinetics was performed in equal share with Dr Mariana Rangel Pereira. I designed the promiscuous activity screen and the cross-inhibition test, selected the substrates and performed the assays jointly with Dr Mariana Rangel Pereira. The analysis, interpretation, presentation and discussion of the results is my own.*

3.2 Introduction

The goal of this chapter is to screen for esterases and lipases in a functional metagenomic screen using droplets. Esterases catalyse the hydrolysis of carboxylic acid esters. Lipases are esterases acting preferentially on long-chain triacylglycerides which are insoluble in water. Amongst other processes, these enzymes are important in the breakdown of biological macromolecules to make them available for cellular metabolism.

Being involved in catabolic processes, the substrate scope of esterases is broad while exhibiting regio- and enantioselectivity, which renders them interesting for industrial applications [113, 114]. Indeed, esterases and lipases are one of the most important enzyme classes in the chemical and biotechnological industries [115]. Lipases find major applications in the modification of fats. One example is the synthesis of structured triacylglycerols such as Betapol which is used as a milk-fat substitute for infants [116]. One large scale application of esterases is the release of ferulic acid from plant cell walls, a compound which is used in the synthesis of vanillin [117]. Esterases are also used to synthesise optically pure compounds and to hydrolyse protecting groups under mild conditions in multi-step organic syntheses [117].

The hydrolysis of esters is achieved by a catalytic triad consisting of Ser-His-Asp, as schematically shown in Figure 3.2 [119]. The aspartate residue favours the tautomeric state in which the histidine enhances the nucleophilic strength of the serine residue. The serine oxygen can thus attack the carbonyl carbon forming the first tetrahedral intermediate. The negative charge of the tetrahedral intermediate is stabilised by two nearby backbone amides, referred to as the oxyanion-hole. The breakdown of the intermediate is facilitated by histi-

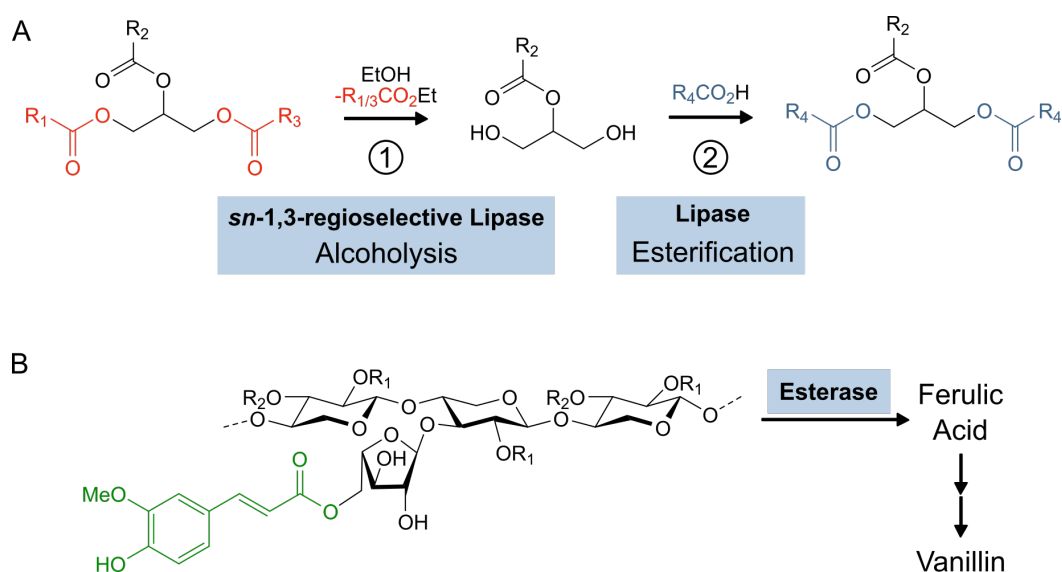


Fig. 3.1 Two important industrial applications of lipases and esterases: the synthesis of structured triacylglycerols (A) and the release of ferulic acid (red part) from plant cell wall polysaccharides (B). Ferulic acid is used to synthesize the important flavouring compound vanillin. Adapted from [116, 118].

dine protonating the leaving alkoxy group. The resulting acyl-enzyme is cleaved by the attack of activated water releasing the acid product via a second tetrahedral intermediate and recovering the free enzyme.

In terms of three-dimensional structure, equivalent catalytic triads have evolved in different protein scaffolds. This constitutes one of the best examples of convergent evolution on the molecular level. The prominent protease trypsin, for instance, belongs to the PA superfamily (25,000 sequences on Pfam¹). Most esterases and lipases belong to the extensive α/β -hydrolase superfamily (270,000 sequences on Pfam).

As enzymes find wider use in industry, there is growing demand for new esterases and lipases to fit a specific purpose [115]. Some corporations offer collections of enzymes that can be screened for their suitability in a certain application [117]. To add to these collections, research has turned to metagenomic screening as a valuable source for new enzymes.

A typical functional metagenomic screen for esterases and lipases consists of transforming the library into a heterologous host organism for expression followed by screening the library on culture plates or in well plate format. The culture plate screen is most commonly used: a growth medium is emulsified with 1% of a triacylglyceride, most commonly tributyrin, which renders the plates opaque. If a colony expresses an active enzyme, it will de-

¹Pfam is a curated protein family database maintained by the European Bioinformatics Institute. Sequences related by homology are grouped into clans (superfamilies) and families. Users can submit protein sequences to assign them to a Pfam family. In this thesis Pfam 31.0 was used [120].

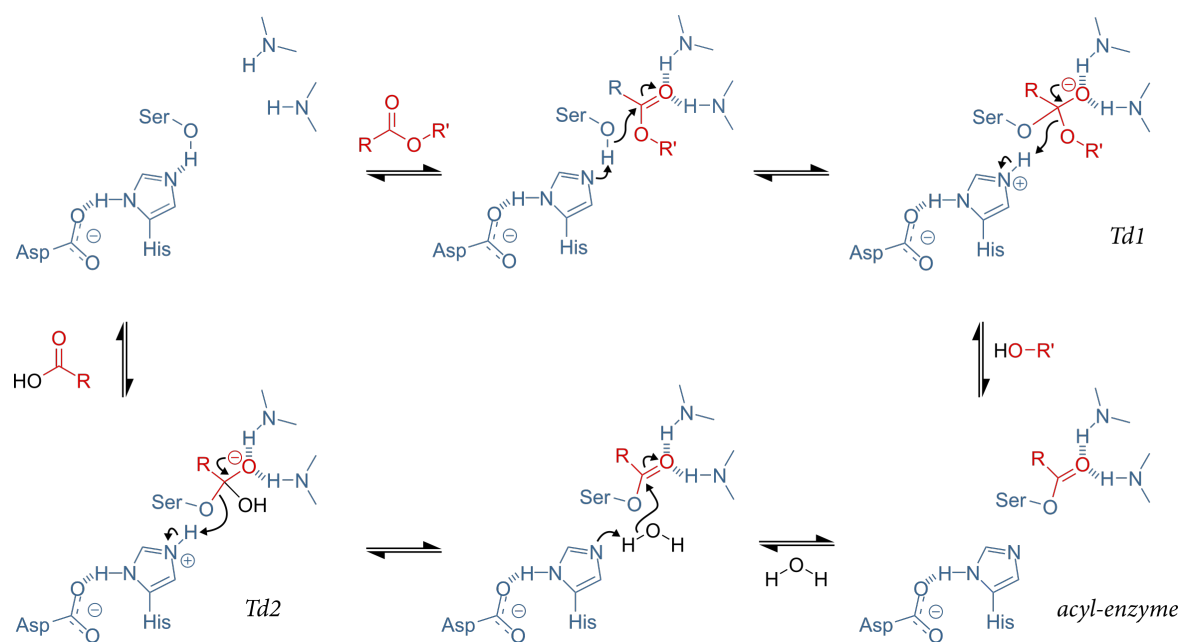


Fig. 3.2 The hydrolysis of esters is catalysed by a catalytic triad consisting of Ser-His-Asp and the so-called oxyanion-hole formed by two backbone amides. The mechanism proceeds via three intermediates. After binding of the substrate, serine attacks the carbonyl carbon atom forming the first tetrahedral intermediate *Td1*, which is stabilised by the oxyanion-hole. The *acyl-enzyme* intermediate formed after the release of the alcohol is attacked by an activated water molecule. The breakdown of the second tetrahedral intermediate *Td2* releases the acid product and recovers the free enzyme. This figure represents the arrangement of residues only schematically. In fact, one of the backbone amides is usually provided by the catalytic serine residue. Also, the catalytic triad only forms upon substrate binding due to the unfavourable dihedral angles the serine residue has to adopt. Figure adapted from [119].

grade the droplets of fat over time and create a clear halo around the colony (Figure 3.3). These colonies are then picked for hit analysis [121, 40, 122–129]. Less frequently used is the well plate format, in which cell lysates are incubated with a chromogenic substrate, for example p-nitrophenyl octanoate [130]. These screening procedures are limited by the number of colonies that can be assayed. Most studies report that 5,000 to 20,000 colonies were tested. Beyond these numbers the plate assays become impractical. Therefore, metagenomic libraries are often under-sampled, *i.e.* fewer clones than a library contains are tested. These practical limits constrain the library size that can be screened, thus the number and diversity of potential hits. Most studies using these screening methods conclude by characterising only one of their hits in more detail, highlighting a single property such as high activity at low temperature, but ignore the potential of any other hit sequences [128].

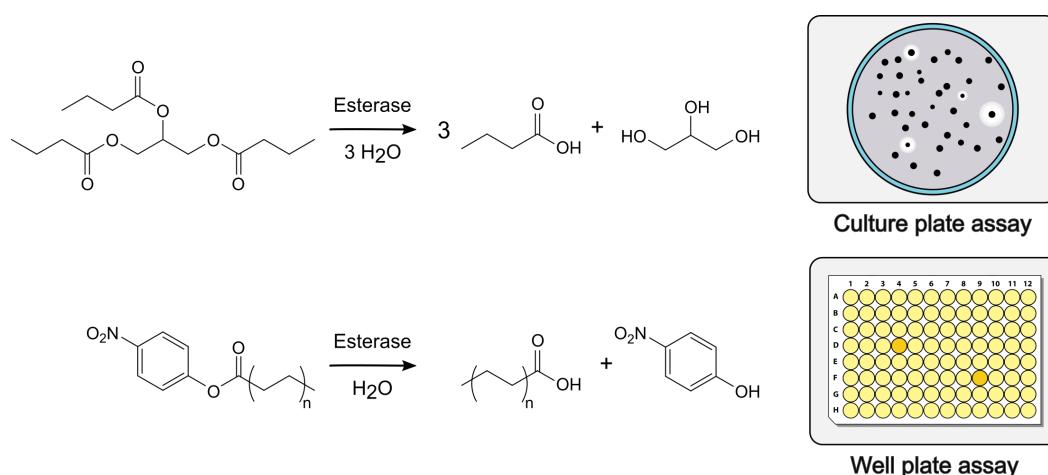


Fig. 3.3 The tributyrin culture plate assay is the most widely used phenotypic screen in functional metagenomics of esterases. Tributyrin is emulsified with the growth medium, rendering an opaque plate. A colony expressing an esterase hydrolyses the compound causing the formation of a halo around it. Less frequently used for screening are well plate assays using p-nitrophenyl esters. The release of p-nitrophenolate can be detected at 405 nm, *i.e.* positive wells turn yellow.

A notable exception is the recently published, largest screening campaign for esterases and lipases to date [127]. In this extremely resource- and time-consuming study, one million clones from 17 distinct metagenomic libraries were screened and 80 new esterases identified by screening over 1,000 tributyrin plates. Four libraries were based on fosmids and screened in arrays of 384 colonies, which amounts to about 150 screened plates. However, 95% of the tested clones were members of λ -phage libraries (lambda ZAP, [131]). These libraries were plated at 1,000 plaques per plate, indicating that at least a further 1,000 plates were screened. Interestingly, 33 of the 80 reported esterases, *i.e.* 40% of the hits, were obtained from the fosmid libraries representing only 5% of the screened clones. Also, almost half of

the hits were contained in only two libraries representing just 7% of the tested clones. After screening and sequencing, the authors select only five hits for further characterisation [127].

It is typical that any one study only characterise a small fraction of the presumed hits in more detail. In two comprehensive reviews Ferrer *et. al* reported, firstly, that in total about 4,000 esterases and lipases had been detected in primary functional metagenomic screens up to 2015, but that, secondly, only 288 were actually validated to be esterases and characterised biochemically, which equates to less than 1 in 10 hits [132, 31]. While they are sometimes tested for their substrate scope, this is exclusively done with carboxylic esters. Any potential promiscuous activities towards other classes of substrates are neglected [121, 40, 122–129].

Authors base their decision on which hit to characterise on homology analysis of their protein sequences. By aligning them against a database such as UniProt the novelty of a sequence is assessed and it is assigned to a protein family. However, the use of this method is increasingly limited because of the vast amount of experimentally uncharacterised sequences available on databases.

However, assigning hits to protein families can be useful to guide further experiments to test their biochemical properties. An important protein family database is Pfam. Another useful resource is ESTHER (ESTerases and α/β -Hydrolase Enzymes and Relatives), a curated database focusing specifically on esterases. This database currently holds 56,000 sequences, which are grouped into Families, Parent Families and Blocks [133]. Another influential classification for bacterial esterases and lipases into eight families was published by Arpigny and Jaeger [134]. Most authors attempt to assign their new enzymes to one of the originally defined families. However, the increasing number of known esterases and lipases has revealed the limitations of this classification. Different authors have extended the number of families to currently eighteen [129, 135].

Taken together, esterases and lipases are needed in industry and metagenomics offers a way of accessing new biocatalysts. The field of functional metagenomic screening for esterases has proliferated many new enzymes using traditional plate assay techniques. However, all previous studies are limited by the number of library members that can be tested. Hence, the establishment of an ultrahigh-throughput screen for esterases and lipases is timely and needed, and is the goal of this chapter. The screening workflow is shown Figure 3.4. The specific aims were:

1. To increase the throughput of esterase discovery by implementing a droplet assay and to screen the SCV library (1.25×10^6 members). Only fluorescence-based droplet assays allow throughputs higher than 10^6 . Fluorescein has been used previously as a reporter molecule to isolate metagenomic hits [53]. Other fluorophores, *e.g.* resorufin and 4-methylumbelliferone, have been shown to readily exchange between droplets,

preventing the detection of enzyme activity after long incubation times [70]. Therefore, fluorescein dicarboxylate was chosen for the initial tests, which can be viewed as a bait substrate for enzyme active sites catalysing hydrolysis.

2. To re-screen the output of the droplet sorting campaign using culture plates containing tributyrin. The formation of a halo around a colony was used to validated hits but the assay may have limited the total number of hits detected because of a lower dynamic range and specific interactions.
3. To sequence the DNA inserts of the isolated plasmids and analyse them. The gene(s) most likely responsible for the observed phenotype were re-cloned to confirm activity and for protein production.
4. To kinetically characterise all enzymes, regardless of the novelty of their sequence, by measuring timecourses with chromogenic p-nitrophenyl carboxylates.
5. To test the concept of fluorescein dicarboxylate acting as a bait by testing the activity of the identified enzymes against a range of compounds that undergo hydrolysis, *e.g.* glycosides and thioesters.

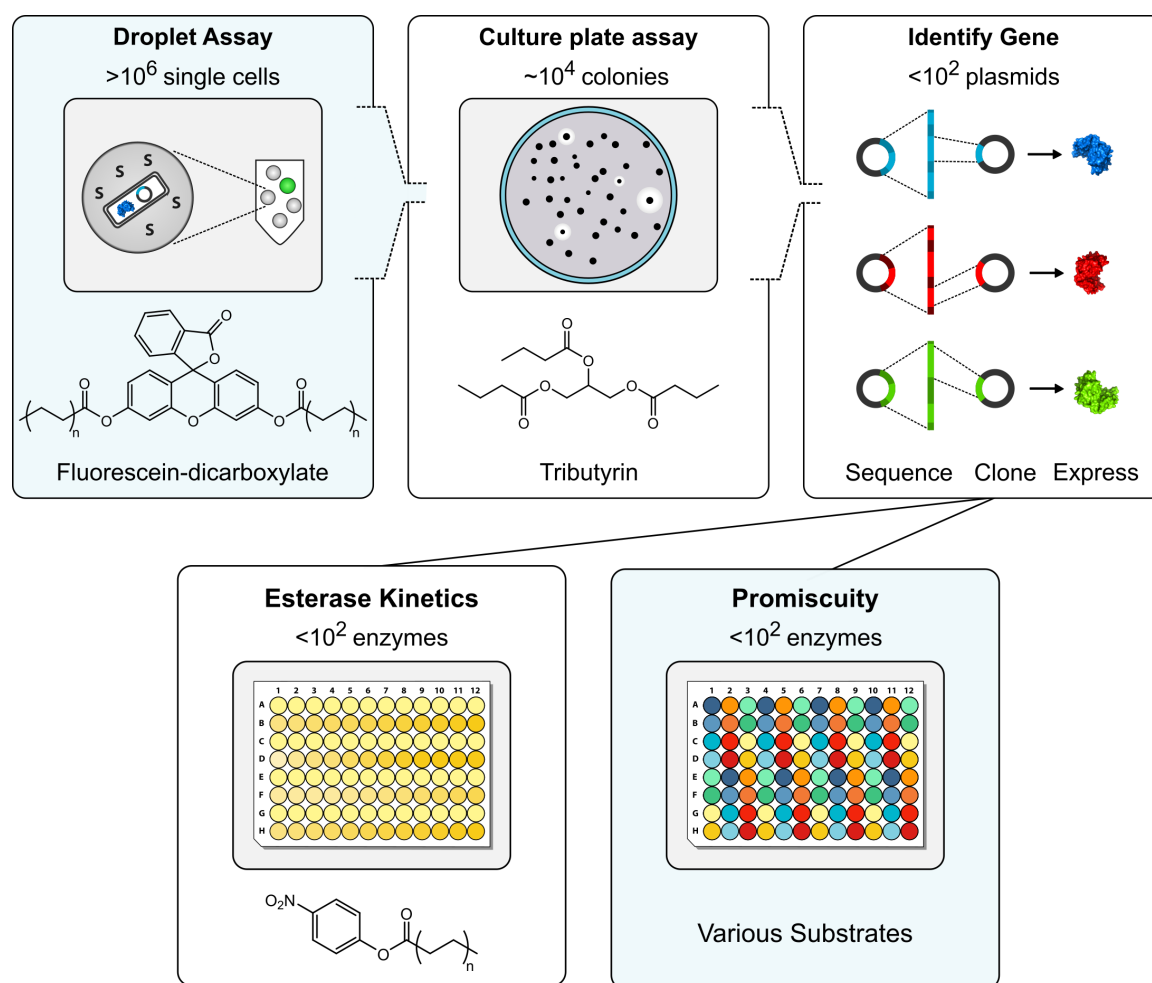


Fig. 3.4 Workflow to isolate esterase gene from a metagenomic library. The approach used in the literature is shown in the white boxes. This will be extended here in two ways (blue boxes): the throughput of the screen will be increased by establishing a droplet assay and the hits will be tested against a range of other substrates.

Table 3.1 Overview of the positive and negative controls used to set up assay conditions for the metagenomic screening.

Protein	k_{cat}/K_m [†] (Ms) ⁻¹	Vector	<i>E. coli</i> strain	Expression
Est30	3.5×10^5	pHAT2	BL21(DE3)	high
Est30Δ	–			
P91	1.0×10^4	pZero2	<i>E. cloni</i> 10G [‡]	low
P35	–			

[†] catalytic efficiency for paranitrophenol acetate.

[‡] commercial cloning strain by Lucigen, derived from *E. coli* K12.

(–) no activity detected.

3.3 Establishing the esterase assay in droplets

Before working on the microfluidic droplet assay, the esterase assay was briefly tested in well plates using cell lysates. The chosen positive and negative controls were Est30 with Est30Δ and P91 with P35, see Table 3.1. Est30 and Est30Δ are a highly active esterase and its inactive mutant [136]. They were overexpressed from pHAT2, a T7 expression system based vector [137]. P91 and P35, an esterase and sulfatase respectively, are hits previously found in the SCV library (Table B.1, [67]). Here, the originally-isolated plasmids with the metagenomic DNA insert were used as genuine library controls, *i.e.* the level of expression could not be controlled but depended on the ability of the *E. coli* host to recognise the foreign DNA [41].

To obtain results in well plate format that are transferable to droplets, the cell lysates of the respective controls were diluted to the concentration that is expected in droplets. The cytosol of a single *E. coli* cell (*ca.* 1 nL) is diluted 2,000 times in a 2 pL droplet. The amount of cell lysate needed to achieve the equivalent dilution in a well plate measurement was calculated from the optical density OD_{600nm} for a cell culture assuming a conversion factor of 5×10^{18} cells/mL for an OD_{600nm} of 1. For example, of a cell culture with OD_{600nm} 3 which was pelleted and resuspended in 100 μL lysis agent, 7.7 μL were added to a 200 μL reaction.

Thus, nine assay conditions were tested as shown in Figures 3.5 and 3.6. Three different fluorescein diester substrates were tested (fluorescein dibutyrate, dihexanoate, and dilaurate) at three different pH values (7, 8, and 9). The goal was to identify assay conditions which allowed overnight incubation to enable even a weak metagenomic hit to hydrolyse sufficient substrate for detection.

Fluorescein dibutyrate is the most similar substrate to the commonly used tributyrin to screen for esterases. However, the difference between the positive and negative metagenomic controls after overnight incubation was smaller than for fluorescein dihexanoate at all pH

values tested. The difference between the controls for fluorescein dilaurate was even smaller. Therefore, fluorescein dihexanoate was selected as the substrate for screening.

The product of the reaction, fluorescein, emits most strongly in its di-anionic form. Given the pK_a of 6.7 of the mono-anion, fluorescein emission plateaus at pH 8 and above [138]. Therefore, pH of 8 was expected to enable the most sensitive detection of product. At this pH, the rate of substrate hydrolysis in the buffer (no cells control) was still low, as seen in Figure 3.6. Given the ratio of positive to negative signal was better at pH 8 than pH 9, pH 8 was selected for screening.

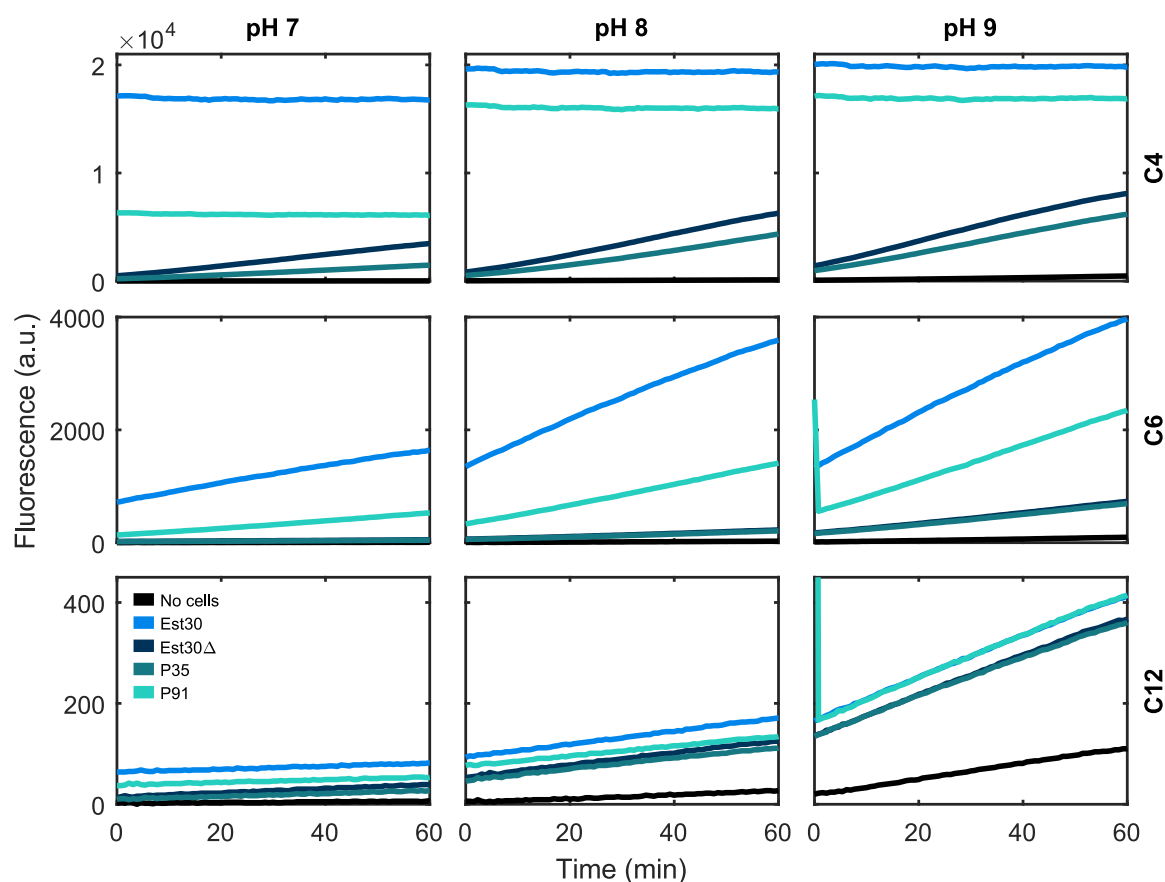


Fig. 3.5 Selection of substrate and pH using nine different assay conditions with positive and negative controls (Table 3.1) over 60 min. From left to right the pH of the buffer (50 mM TrisHCl, 100 mM NaCl) is 7, 8, and 9 respectively. From top to bottom the carbon chain length of the fluorescein diester (20 μ M) increases from 4 (dibutyrate), to 6 (dihexanoate), and to 12 (dilaurate). Fluorescein dibutyrate was fully hydrolysed by the positive controls before the measurement commenced. However, high rates of hydrolysis in the negative controls were also observed on this timescale. Fluorescence was excited at 480 nm and measured at 520 nm.

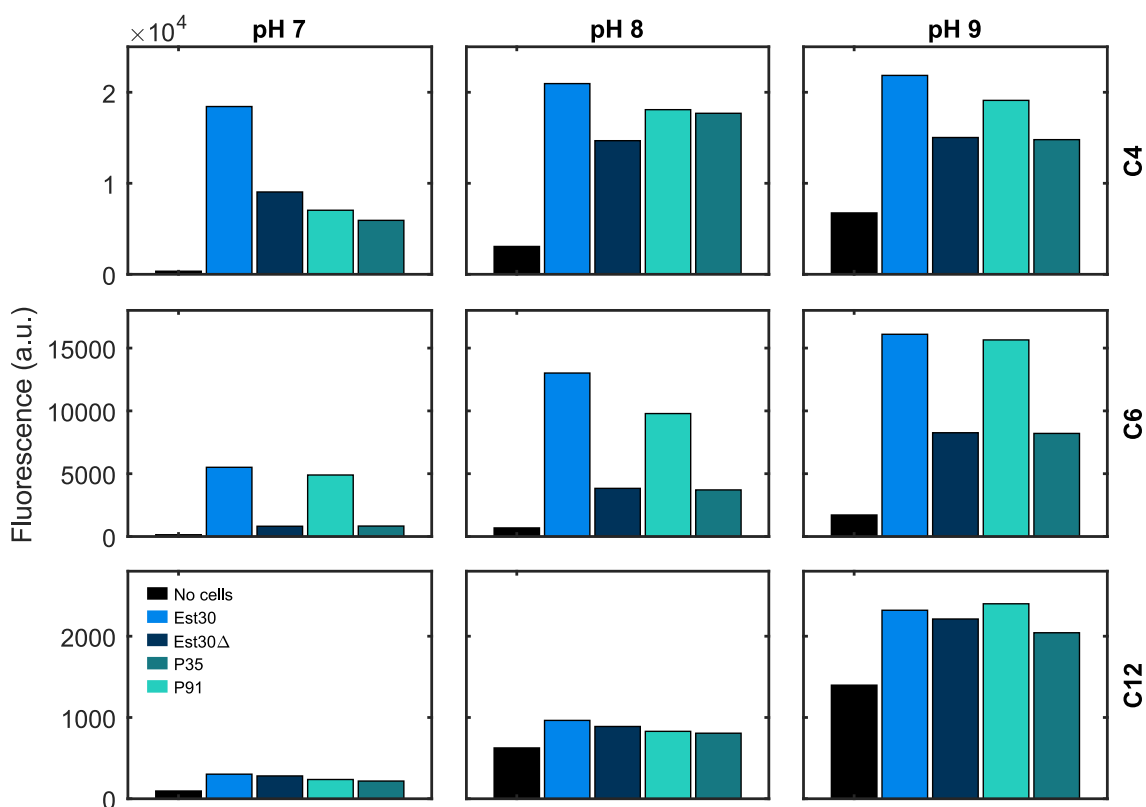


Fig. 3.6 Fluorescence reached after overnight incubation (18h) of the nine reactions shown in Figure 3.5. Long incubation times are beneficial for functional metagenomic screening to isolate weak or weakly expressed enzymes. However, this is only possible if the background hydrolysis of the substrate is sufficiently low. The difference between positive and negative controls was largest for fluorescein dihexanoate, while pH 8 is offered the best trade-off between high positive and low negative signal. Fluorescence was excited at 480 nm and measured at 520 nm.

Using the chosen conditions, 15 μm droplets were generated containing cells that had overexpressed Est30 and Est30 Δ . The cell cultures were premixed at a positive to negative ratio of 1:100 and encapsulated at an occupancy λ of 0.1. Therefore, 0.1% of the total droplet population, or 1 in 1,000 droplets, was expected to develop strong fluorescence above background. Indeed, fluorescence microscopy 3h after droplet generation showed bright fluorescent droplets at about the expected occupancy. The micrograph in Figure 3.7 contains 8 fluorescent droplets out of about 7,000 (0.11%) in close agreement with the calculated percentage.

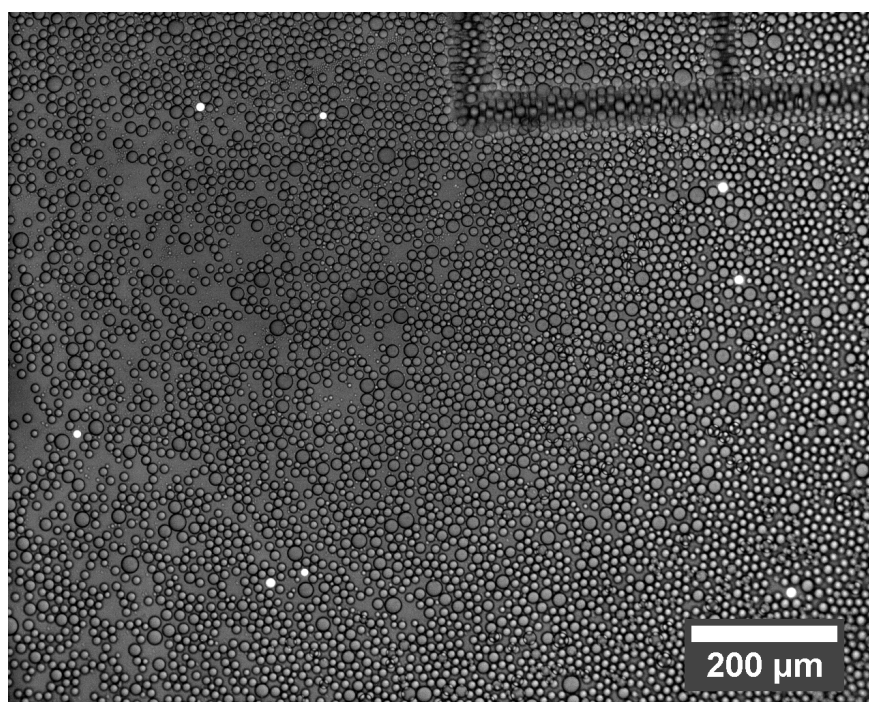


Fig. 3.7 Fluorescence micrographs of droplets containing cells which expressed Est30 and Est30 Δ at 1:100 and with an occupancy λ of 0.1. The brightly fluorescent droplets (white) were observed at the expected frequency for the positive control Est30, indicating that the esterase assay worked in droplets and had a good signal to noise ratio.

The fluorescence distribution of these droplets was determined by FADS (Chapter 2). The histograms after 3h and 18h incubation are shown in Figure 3.8. The population above 400 a.u., which represents saturation of the detector in this configuration, is 0.12% and 0.16% respectively, and therefore likely to be the signal stemming from Est30 activity. The shoulder at 100 a.u. appearing after overnight incubation could be explained by the background activity of Est30 Δ seeing as it corresponds to about 10% of the total population.

Encouraged by these results, droplets were created containing the P91 and P35 controls. However, when mixed at a ratio of 1:100, no distinct population of high fluorescence was

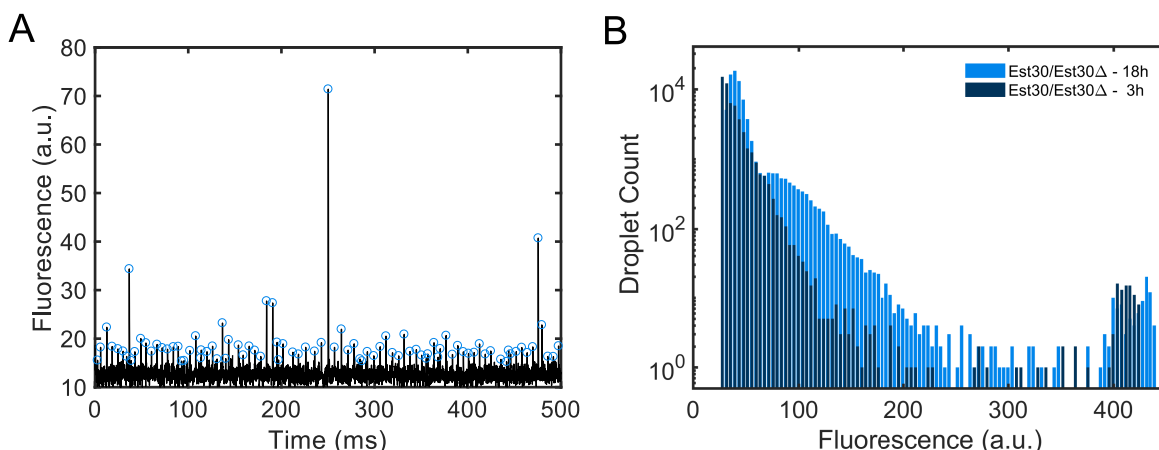


Fig. 3.8 Esterase assay with fluorescein dihexanoate in droplets using Est30 and Est30 Δ expressing cells in a ratio of 1:100 and at occupancy λ of 0.1. *A*: Timeseries of the fluorescence signal of analysed droplets. Each peak is caused by a droplet passing through the laser beam. The droplet histogram in panel *B* was generated in real-time by recording the amplitude (blue circles) of all the peaks. The fluorescence of negative droplets was caused by partial background hydrolysis of the substrate, no additional fluorophore was added. *B*: Distribution of droplet fluorescence after 3h and 18h of incubation. The percentage of droplets with a signal above 400 a.u. is 0.12 and 0.16% respectively, in good agreement with the expected 0.1% of positive events.

observed (data not shown). Therefore, separate droplet samples were generated for the two controls. As shown in Figure 3.9, after overnight incubation the frequency of droplets with fluorescence above 100 a.u. was much higher in the P91 sample than the P35 sample. However, the fluorescence values were distributed over the whole signal range, indicating strong variation in the amount of protein produced from cell to cell. This increased variation in the P91 sample compared to Est30 Δ sample is rationalised by the difference in the expression constructs. As explained above, the expression of the former relied on spurious transcription by the *E. coli* host from the metagenomic DNA, while the expression of the latter was under tight control of a T7 promoter in the pHAT vector. Interestingly, the background in the P35 sample did not increase as much as in the Est30 Δ case, likely due to the different *E. coli* strain used.

Finally, the SCV library was transformed and encapsulated into droplets using the chosen conditions (Figure 3.9). The fluorescence distribution revealed rare events above 200 a.u. which were indicative of the detection of esterase activity in the actual metagenomic library. Given this result, a droplet sorting campaign was performed, to isolate the library members responsible for the signal, as reported in the next section.

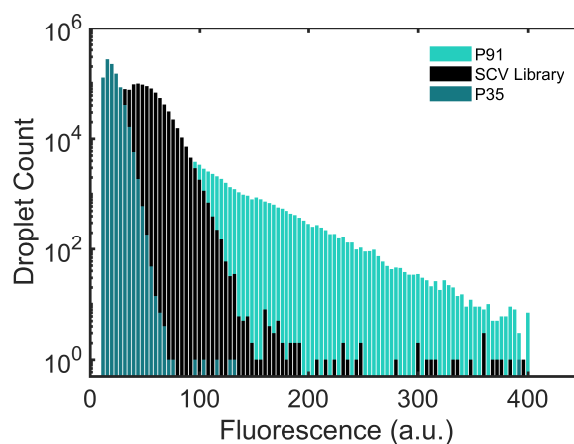


Fig. 3.9 Distribution of droplet fluorescence in three distinct droplet samples ($\lambda=0.35$). Fluorescence above 100 a.u. was much more frequent in the P91 sample compared to the P35 sample as expected. However, no defined population separate from the background was discernible. The sample with cells expressing the metagenomic SCV library showed events with fluorescence above 200 a.u., indicative of the presence of esterase activity due to a library member.

3.4 Screening of the SCV Library

Taking the insights gained in the previous section, the workflow shown in Figure 3.10 was devised to screen the metagenomic library for esterases. The principle task was to isolate individual plasmids conferring esterase activity to cells from a pool of over 1×10^6 plasmids which was the SCV library. To achieve this, first, the plasmid library was transformed and the transformed cells incubated on plates for two days to allow time for the expression of protein. The cells were then encapsulated with substrate and lysed, incubated for the reaction to occur, and finally sorted. The DNA was recovered from the collected droplets and re-transformed into cells to amplify it. To distinguish real hits from false positives, the transformed cells were plated onto tributyrin containing plates, *i.e.* the plates were opaque. They were checked for the formation of a clear halo around the colonies, indicating esterase activity. Both colonies with halos and, as a control, with no halos were picked and grown as separate cultures to recover the plasmid for sequencing and downstream characterisation.

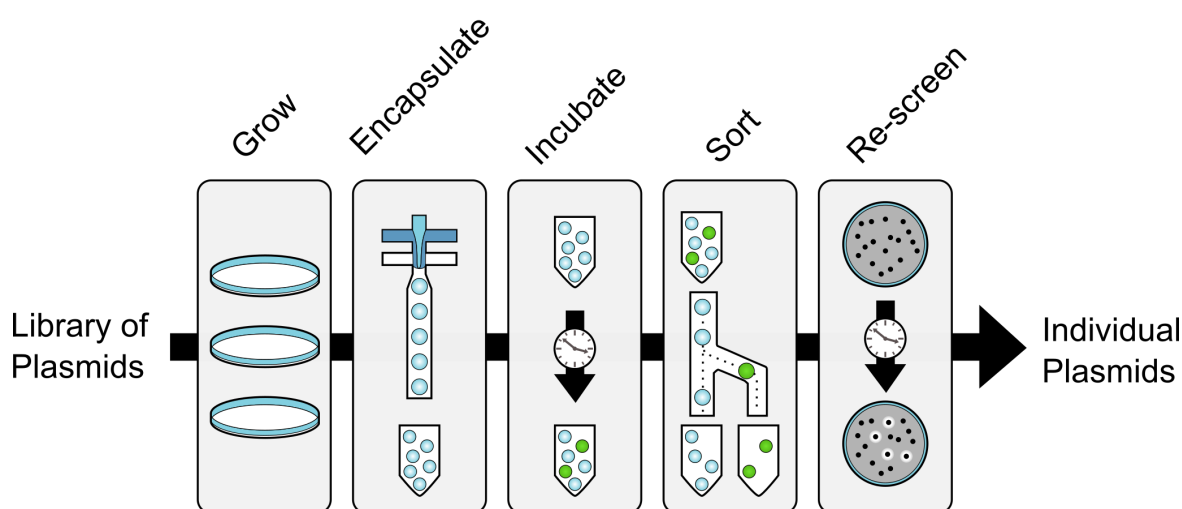


Fig. 3.10 The figure shows schematically how the metagenomic SCV library was sorted to recover individual plasmids conferring esterase activity to cells. Section 3.4 describes the procedure.

3.4.1 Droplet screening and hit recovery

The library was screened in two separate screening campaigns (campaign I and II) each following the workflow above. In both cases the droplets were generated with an occupancy λ of 0.35. Batches of these droplets were sorted up to three days after their generation. The histograms of all droplet sorts are shown in Figure 3.11.

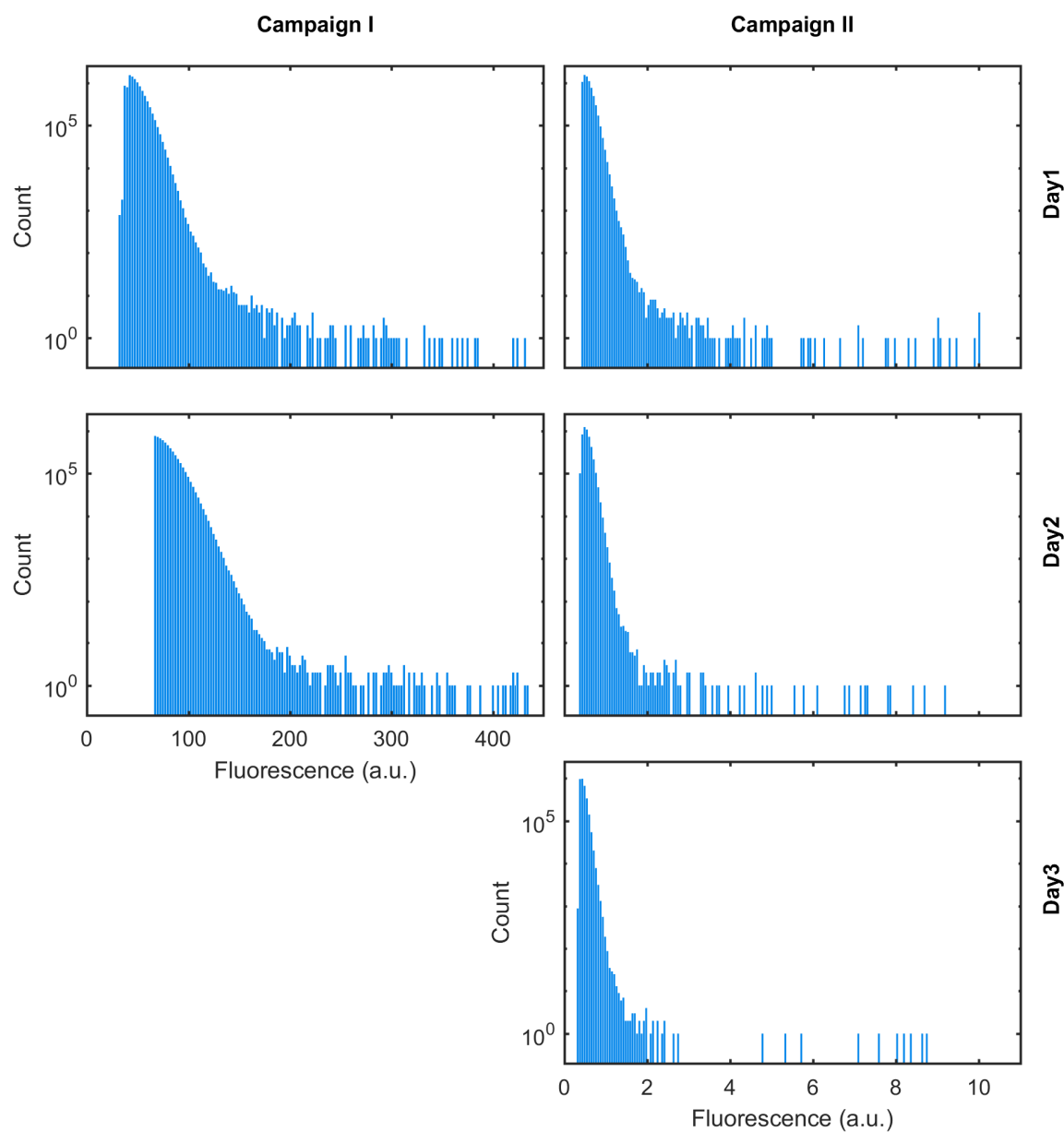


Fig. 3.11 Fluorescence histograms of the droplet sorting of the SCV library using fluorescein dihexanoate. Two separate screening campaigns were performed. During each campaign droplets were sorted over several days to assess if prolonged incubation influenced the hit rate. The difference in the scales of fluorescence between campaign I and campaign II is due to the use of different fluorescence detectors, refer to Section 2.4.1 for details.

Table 3.2 Summary of the two SCV library sorting campaigns.

Campaign	Day	Droplets			Colonies		Unique Hits
		Sorted /10 ⁶	Collected	Hit Rate	Total /10 ³	Halos	
I	1	10.0	126	3.5×10^{-5}	6	17	
	2	5.7	110	5.5×10^{-5}	10	17	
	Total	15.7	236	4.3×10^{-5}	16	34	11
II	1	7.0	65	2.6×10^{-5}			
	2	4.8	104	6.2×10^{-5}			
	3	3.1	77	7.1×10^{-5}			
	Total	15.0	246	4.7×10^{-5}	ND	ND	6

ND, not determined.

The key numbers of the two campaigns are summarised in the Table 3.2. In campaign I, 15.7×10^6 droplets were sorted over two days, *i.e.* an estimated 5.5×10^6 single cells, which equates to about 4.4-fold over-sampling of the library. The 126 and 110 most fluorescent droplets were collected on the first and second day, respectively. In campaign II, 15.0×10^6 droplets were sorted over three days with a focus on longer incubation time. A similar number of droplets was collected for recovery. An apparent hit rate was calculated based on the number of droplets collected per number of clones screened. It was in the range of 10^{-5} , typical for functional metagenomics, [31], and increased with longer incubation times, indicating that the reaction continued to accumulate product up to three days.

Next, the DNA of the collected droplets was recovered and re-transformed separately for each sorting day. Campaign I yielded about 16,000 colonies in total. The plates were monitored up to two weeks and 34 colonies with clear halos were picked for downstream analysis, see Figure 3.12 for representative examples. Thus, the hit rate in the culture plates was in the range of 10^{-3} suggesting an average enrichment of 10^2 from droplets to plates. Upon sequencing it was found that several halo-forming colonies harboured the same library plasmid. The number of unique hits was 11 in campaign I, *i.e.* on average every hit was found three times (a more detailed analysis follows in the next section). A high number of false positives, *i.e.* colonies that did not gain a halo with time, was not unexpected as it had been observed by Colin *et al.* in their metagenomic screening campaigns [67, 84]. The naïve library plated at a similar colony density did not yield any colonies with clear zones around them, confirming that the droplet sorting was essential to recover hits from this library.

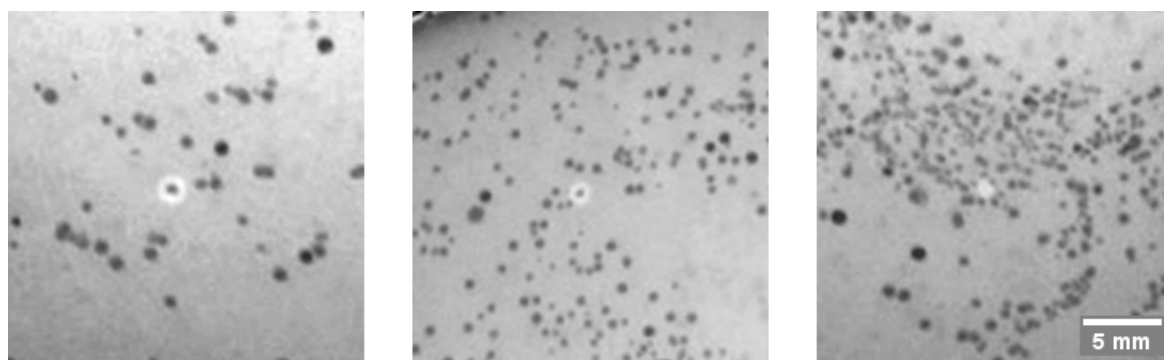


Fig. 3.12 Photos of tributyrin-containing agar plates with *E. coli* colonies recovered post-sorting. At the centre, each photo shows a colony surrounded by a clear zone, or halo, indicating the presence of an esterase.

The clones selected during campaign I were re-grown in a 96-well plate in triplicates. Their diluted lysates were incubated with fluorescein dihexanoate to test their activity (Figure 3.13). The figure shows the activity after 45 min of incubation. All tested clones showed activity in the range between the negative and positive controls. The highest activity was observed for N18. After having confirmed the activity of the hits on the substrate used in the droplet screening, their plasmids were isolated for sequencing.

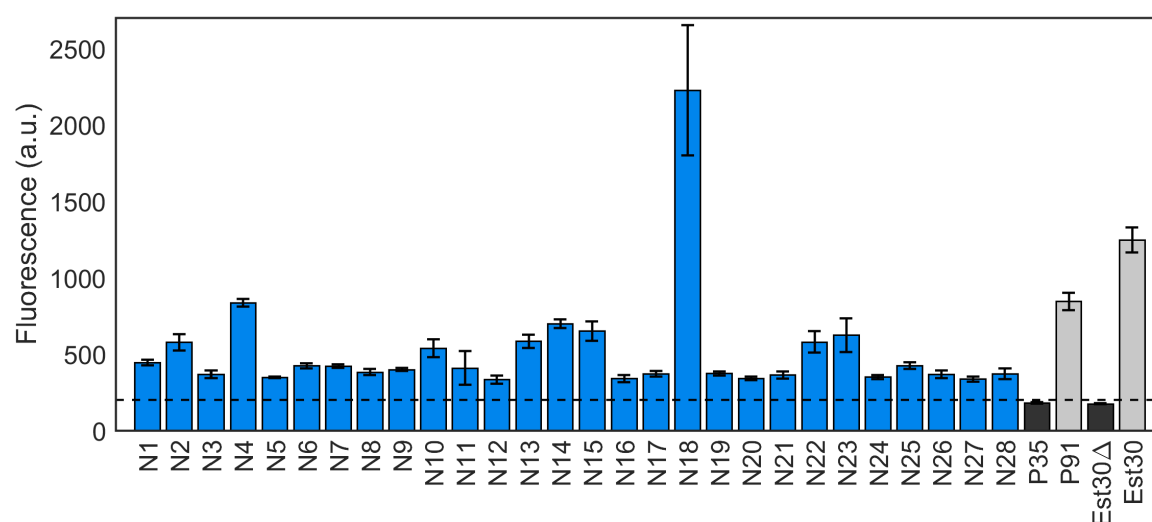


Fig. 3.13 Activity of cell lysates incubated with 20 μ M fluorescein dihexanoate was above background (dashed line) after 45 min of incubation. The lysates were diluted to the extent expected in droplets (2×10^3) in 50 mM TrisHCl pH 8.0, 100 mM NaCl. Negative controls: P35 and Est30 Δ , positive controls: P91 and Est30. Dashed lines: three standard deviations above P35 fluorescence (the Est30 Δ value was lower). Error bars are the standard deviation of triplicate measurements.

Table 3.3 This table lists the selected positive clones, their DNA insert size, and how often they were found during each screening campaign.

Clone	Restriction Site [†]	Insert size (kbp)	Campaign I	Campaign II	Total
N1	EcoRI	4.2	9	1	10
N2	EcoRV	5.3	1	1	2
N4	EcoRV	3.9	3	8	11
N7	EcoRV	4.1	2	2	4
N11	EcoRI	2.4	2		2
N13	EcoRV	3.1	2		2
N16	EcoRI	2.9	3		3
N18	EcoRV	3.4	1	2	3
N20	EcoRI	2.5	2		2
N26	EcoRV	1.2	1		1
N33	EcoRV	3.5	1		1
RR11	EcoRI	2.2		1	1

[†] The cow rumen sub-library of SCV was constructed using the EcoRI restriction site, whereas all others were constructed using EcoRV.

3.4.2 Sequence analysis of the selected clones

All clones selected in the two screening campaigns were sequenced using the standard M13 forward and reverse primers flanking the DNA inserts on the library vector pZero2 (for vector map see Appendix Figure B.1). The initial sequences revealed that some of the halo-forming clones were identical. In total, 12 unique clones were found, as listed in Table 3.3. Nine hits were found more than once. The clones N1 and N4 were found most frequently: 10 and 11 times each. The clones N26, N33 and RR11 were only found once. The multiple recovery of several hits was indicative that both the library coverage and the DNA recovery were efficient at isolating these hits. The restriction site used to insert environmental DNA into pZero2 was EcoRV for most of the sub-libraries contained within the SCV library [84]. Only in the case of the cow rumen library, which contains 10% of the screened clones, was EcoRI used. Five of the twelve could therefore be assigned to the cow rumen library, indicating that the hit rate for this library was 6× higher than the average of the other libraries.

The only clone whose forward and reverse sequences could be aligned to obtain the full DNA insert after the first round of sequencing was N26, which had the smallest insert at 1.2 kbp. The other clones required up to three rounds of sequencing by primer walking, *i.e.* sequencing using custom primers designed to bind near the ends of the known insert

sequence. All DNA inserts could thus be sequenced. The largest insert had a size of 5.3 kbp, while the average was 3.3 kbp.

Next, the open reading frames (ORFs) responsible for the detected esterase activity needed to be identified. As a first assessment, the predicted protein sequences of all the ORFs larger than 300 nucleotides were searched against the NCBI non-redundant protein database using Basic Local Alignment Search Tool (BLAST). All DNA inserts contained at least one predicted protein that had detectable sequence similarity with a sequence which was annotated as an esterase, lipase or α/β -hydrolase in the database. In total, 13 unique protein sequences were found to be the likely esterase candidates. The location of each of their respective ORFs is shown in Figure 3.14. The protein sequences identified in the inserts of clones N18 and N33 were identical. It was revealed that these two inserts could be aligned to each other, *i.e.* they must have derived from the same stretch of environmental DNA used to construct the library.

When searched for in October 2016, the sequence identities on the DNA level of the top hit for N1ORF4 and N1ORF5 had been 34% and 32% respectively. Since then, new sequences were deposited on the database and the identity with the top hits is now 88% and 90%. With similar effect, the top hit changed for 10 out of the 13 unique protein sequences. As of September 2018, 7 sequences had identities equal or larger than 90% with a deposited sequence, refer to Table B.2 for a detailed list.

Nonetheless, all of the sequences constituted new, uncharacterised proteins. The top hits listed by BLAST for each predicted protein from this study were predicted proteins themselves with no prior experimental evidence underlying the annotation on the database. Only in the case of RR11ORF2 was the top hit an experimentally verified esterase (85% sequence identity): EstGK1, which was found in a traditional metagenomic screening campaign in 2010 [130]. However, the highest sequence identities with the next characterised protein for all other hits ranges from not detectable to a maximum of 68% as shown in Table 3.4. That is, it was of interest and necessary to clone each of the identified ORFs to prove that they encoded for an esterase or lipase.

Prior to this, the protein sequences were further analysed. All hits were searched for matching family entries in the Pfam and ESTHER protein family databases. On Pfam, 10 out of the 13 unique protein sequences were predicted to contain a single domain belonging to the α/β -hydrolase clan (CL0028), which contains most known esterase sequences (Figure 3.15). They were assigned to a number of esterase and lipase families within this clan. Interestingly, 4 of the 13 hits were assigned to small protein families together consisting of less than 4,000 sequences, whereas the remaining 7 were part of families with over 170,000 members. That is, one third of the hits were from families constituting only 2% of the sequence space of inter-

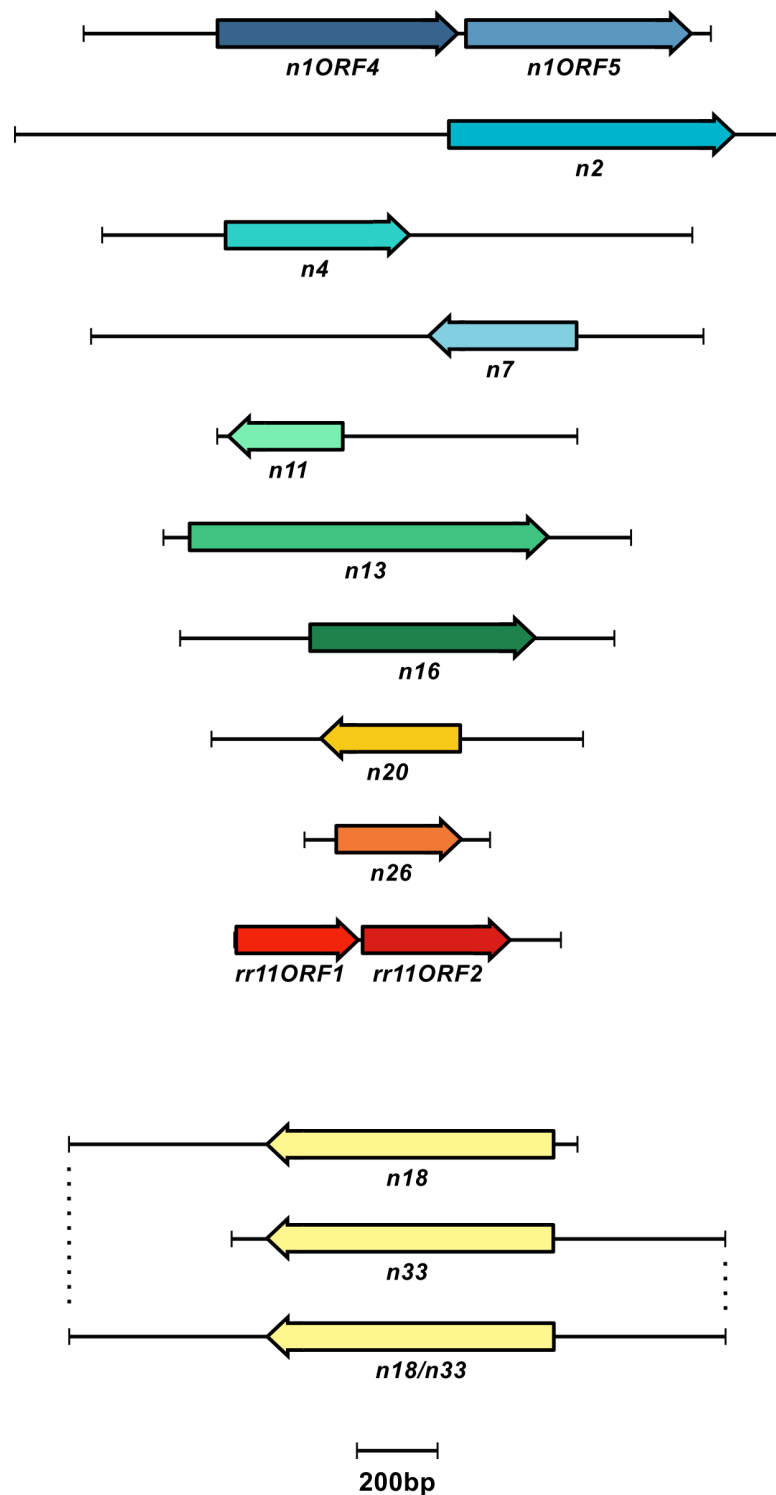


Fig. 3.14 This figure shows the ORFs related to esterase genes identified in the sequenced clones. The orientation is 5' to 3' end with respect to the vector pZero2, the length of the DNA inserts and ORFs are to scale. The DNA inserts of clones N18 and N33 contained an identical ORF and could be aligned to form a continuous sequence.

Table 3.4 This table lists the most closely related enzyme (by sequence) for each hit which was biochemically confirmed to have esterolytic activity.

Hit	Enzyme [†]	Query Cover	Identity	Reference
N1ORF4	PNBCE	90%	34%	[139]
N1ORF5	PNBCE	94%	32%	[139]
N2	EstA	99%	65%	[140]
N4	–	–	–	–
N7	N-Acetyl Hydrolase	65%	33%	[141]
N11	XynC	96%	41%	[142]
N13	Extracellular Lipase	61%	64%	[143]
N16	PNBCE	92%	33%	[139]
N18	EstA	98%	59%	[140]
N20	EstGK1	96%	68%	[130]
N26	XylF	87%	29%	[144]
RR11ORF1	Arylesterase	76%	32%	[145]
RR11ORF2	EstGK1	98%	85%	[130]

[†] Top BLASTp hits against the non-redundant protein or UniProt databases associated with publications that prove the activity of the enzyme, as of September 2018.

– no characterised protein with homology to N4 could be found.

est. Notable are N20 and RR11ORF2 which were assigned to DUF3089, a family which was only recently discovered with few characterised members [130, 146, 147]. Hits N2 and N18 were each predicted to contain two domains. One was assigned to the SGNH-hydrolase clan (CL0264), which contains the lipases of the GDSSL family; the second was a transmembrane β -barrel domain (CL0193). These two hits appear to be trans-membrane proteins with an extracellular lipase domain similar to EstA with which they show 65 and 59% sequence identity respectively [140, 148].

To gain further insight, the sequences of the hits were also submitted to the ESTHER database, which specialises in the α/β -hydrolase superfamily and is manually curated [133]. The classifications between the two databases differ, with ESTHER containing more functionally characterised entries. As expected, N2, N13, and N18 could not be assigned to a family in this database since they are not members of the α/β -hydrolase superfamily. Noteworthy assignments are N11 and N26. The hit N11 was assigned to the family of feruloyl esterases (within the Antigen85c family) for which there were only 300 known members. This family of esterases is of commercial importance as mentioned above [117]. The hit N26 was assigned to the epoxide-hydrolase like family, specifically the C-C bond hydrolases for which also only 300 sequences were known. These enzymes hydrolyse conjugated 1,5-diketones [149, 150], they are catalytically promiscuous [151], and they have previously been found in a metagenomic screening [152]. Therefore, the functional screen isolated hits from across the spectrum of known esterase families. There was a notably high proportion of hits from small families, *i.e.* they were from families of rarely observed enzymes, which have become accessible thanks to metagenomics at ultrahigh-throughput.

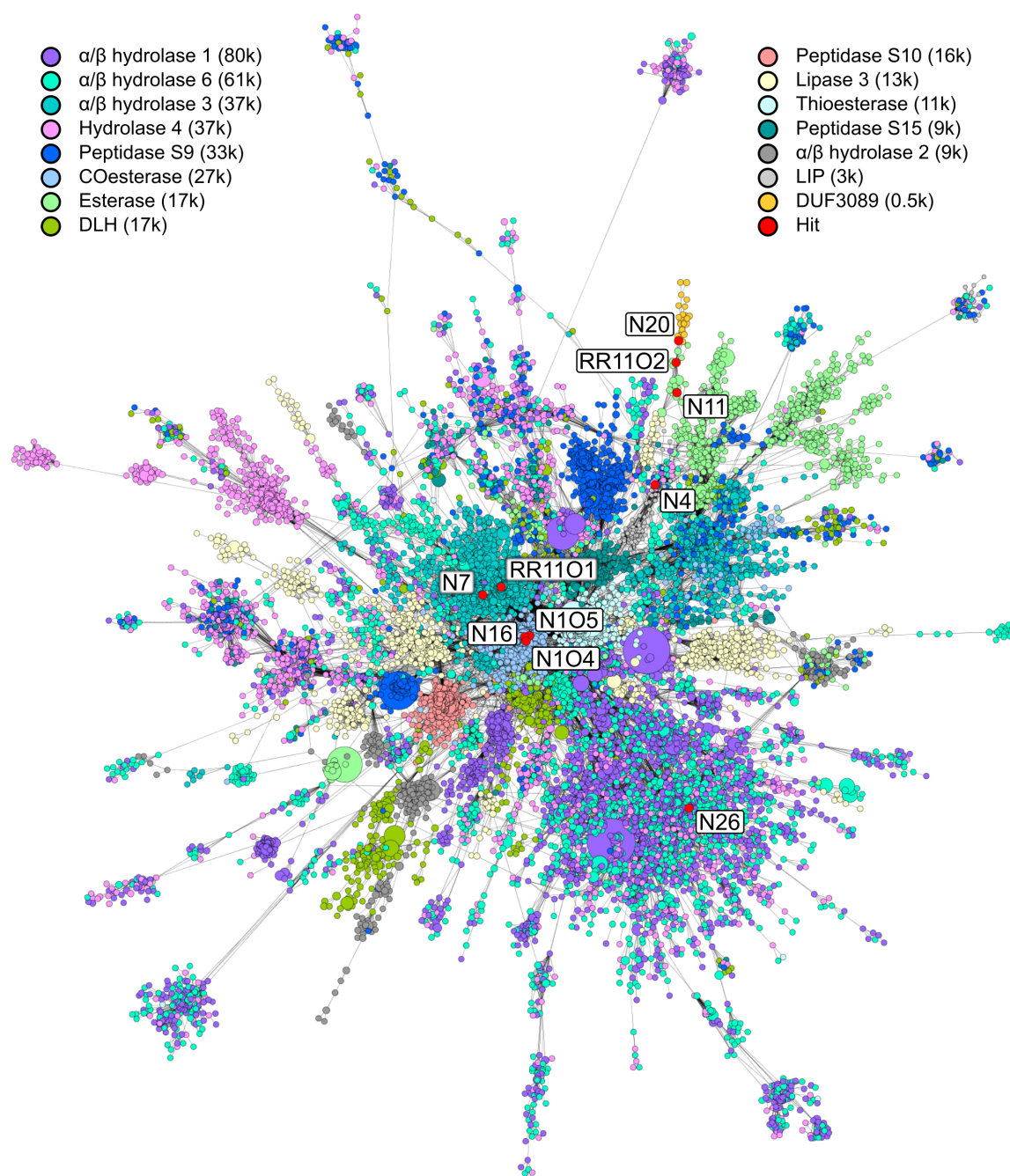


Fig. 3.15 Sequence similarity network of the α/β -hydrolase superfamily (Pfam CL0028) containing 10 of the 13 hits. Sequence similarity networks are an alternative to phylogenetic trees showing how large numbers of sequences are related [153]. Hits N4, N20 and RR11O2 are located in clusters representing less than 1% of the displayed sequence space. N11 was assigned to the Esterase family, but is located in a small sub-cluster neighbouring the DUF3089 family. The network contains 370,000 sequences clustered into 20,000 nodes representing 80% of CL0028. The node size is proportional to the number of sequences represented (range 1 to 2600, median 4). Connecting lines represent alignments with e-values $<10^{-20}$, the median alignment length was 312. Networks for N2, N13 and N18 shown in Appendix Figures B.3 and B.4.

Table 3.5 Protein families that each hit belongs to in the Pfam and ESTHER databases respectively.

Name	Pfam [†]	ESTHER [‡]		
		Clan	Family	Subfamily
N1ORF4	AB_hydrolase	COesterase	Carb_B_Bacteria	
N1ORF5	AB_hydrolase	COesterase	Carb_B_Bacteria	
N2	SGNH_hydrolase	Lipase_GDSL	-	
	MBB	Autotransporter	-	
N4	AB_hydrolase	LIP	Fungal-Bact_LIP	
N7	AB_hydrolase	Abhydrolase_3	GTASGmotif	
N11	AB_hydrolase	Esterase	Antigen85c	A85-Feruloyl-Esterase or A85-EsteraseD-FGH
N13	-	Lipase_bact_N	-	
N16	AB_hydrolase	COesterase	Carb_B_Bacteria	
N18*	SGNH_hydrolase	Lipase_GDSL	-	
	MBB	Autotransporter	-	
N20	AB_hydrolase	DUF3089	Duf_3089	
N26	AB_hydrolase	Abhydrolase_1	Epoxide-hydrolase_like	Carbon- carbon_bond_hydrolase
RR11ORF1	AB_hydrolase	Abhydrolase_3	Hormone- sensitive_lipase_like	
RR11ORF2	AB_hydrolase	DUF3089	Duf_3089	

[†] The Pfam database was searched for each sequence using HHMERscan with e-value 1.0 [19].

[‡] The ESTHER database was searched for each sequence using its HMMER interface with e-value 0.05 [133]. The reported families are the respective top hits for each search.

* The ORF of interest on N18 was identical to the ORF in N33.

(-) No assignment.

Most of the hits were assigned to the α/β -hydrolase superfamily of proteins. The α/β -hydrolase superfamily was first defined by Ollis *et al.* in 1992 [154]. As shown in Figure 3.16, the canonical fold consists of a β -sheet surrounded by α -helices. The β -sheet consists of eight strands of which strand 2 is antiparallel to the others. Strands 3 to 8 are connected by α -helices. The catalytic residues are located on loops between strand 5 and helix C (Ser), strand 7 and helix E (Asp), and near the C-terminus (His). Adaptation to different substrates by individual family members is achieved by differing cap domains above the catalytic triad. The cap domains are formed by peptides that extend from the C-terminal ends of strands 4 and 6-8. Originally defined using five enzymes, the α/β -hydrolase fold has grown into a large superfamily of proteins catalysing numerous chemical reactions [155]. As said, many of them are esterases/lipases but also peptidases. Other hydrolytic enzymes found in this superfamily are epoxide hydrolases, haloalkane dehalogenases, enolactotases, and C-C bond hydrolases. The functional versatility of this fold goes even beyond that: Rauwerdink and Kazlauskas counted the catalysis of 17 distinct reaction mechanisms including oxidoreductive and lyase mechanisms [119]. Taking this together, it was likely that the hits had a similar tertiary structure and that the hydrolytic activity observed for the identified hits was brought about by an active site with a catalytic triad. Furthermore, due to the functional diversity of this fold, promiscuous activities of these enzymes were likely to exist.

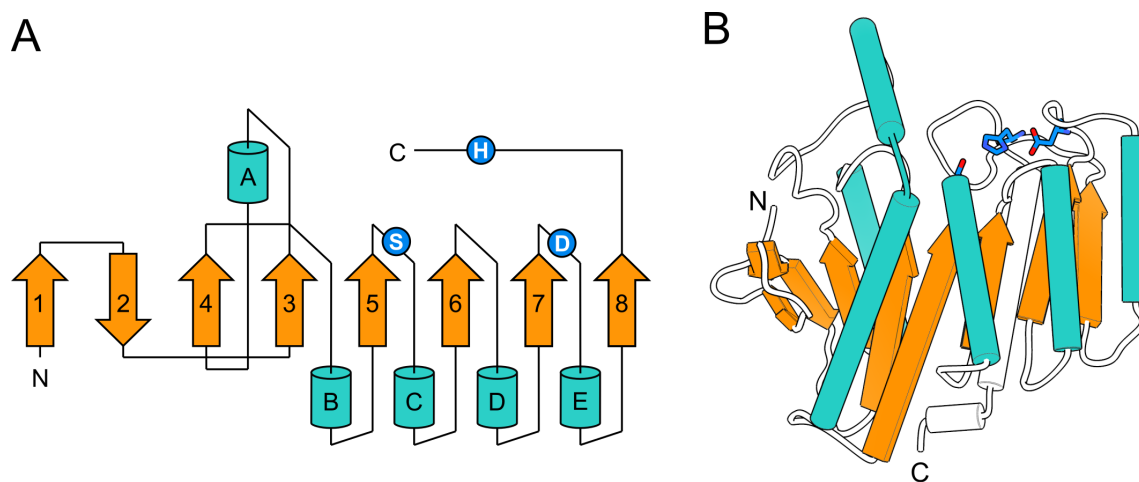


Fig. 3.16 This figure shows the canonical secondary and tertiary structures of the α/β -hydrolase fold *A*: The secondary structure shows how strands 4 to 8 of the central β -sheet are connected to each other by five intervening α -helices. The residues forming the catalytic triad are located on loops as indicated by the blue circles. *B*: The tertiary structure of diene lactone hydrolase (PDB: 4P93), a near canonical α/β -hydrolase, shows the relative orientation of the catalytic residues in the top part of the structure. Accommodation of different substrates in family members of this fold is achieved by insertions into the loops connecting the β -sheets and α -helices. Figure adapted from [155].

In summary, each DNA insert contained an ORF which was predicted to be linked to ester hydrolysis. Further analysis of these ORFs showed that they all represented new, previously uncharacterised sequences. The sequences were assigned to a diverse range of esterolytic protein families some of which have interesting catalytic properties. Next, it needed to be experimentally verified that the identified ORFs were indeed responsible for the esterase activity of the respective clone.

3.5 Quantitative analysis of the newly identified esterases

The predicted ORFs were cloned into the vector pHAT for gene expression in *E. coli* BL21(DE3) [137]. Using the SpeI and HindIII cloning sites added an N-terminal 6xHis-tag to all proteins, which enabled purification using immobilized metal ion affinity chromatography (IMAC). For all hits, the entire gene was cloned, except in the case of N2, where only the lipase domain was cloned (residues 25-358), and N4, where a signal peptide was removed (first 19 residues), see Appendix Table B.3 and Figure B.6. For hit N18, no expression construct was found to yield sufficient amounts of protein for Michaelis-Menten kinetics.

3.5.1 Esterase kinetics with p-nitrophenyl carboxylates

The Michaelis-Menten kinetic parameters of the purified proteins were determined for p-nitrophenyl carboxylates. The substrate concentration was varied and the initial rates of reaction measured. The Michaelis-Menten equation was then used to determine the kinetic parameters [156]:

$$v_i = \frac{v_{\max}[S]}{K_m + [S]}; \quad v_{\max} = k_{\text{cat}}[E_0] \quad (3.1)$$

With v_i and v_{\max} being the initial rate and maximal rate, respectively; $[S]$ the substrate concentration; K_m the Michaelis constant, which is a measure of the enzyme's affinity to a substrate; k_{cat} the turnover number, which is maximum number of substrate conversions per second per enzyme active site; and $[E_0]$ the enzyme concentration. The catalytic efficiency k_{cat}/K_m is generally used to compare different enzymes. Figure 3.17A shows a typical measurement for hit N7 with substrate p-nitrophenyl acetate.

The chain length of the p-nitrophenyl carboxylate was varied from two to sixteen carbon atoms. Again, typical results are shown for N7 in Figure 3.17B. For N7, the highest catalytic activity was observed for p-nitrophenyl hexanoate, followed by similar activities for p-nitrophenyl acetate and butyrate. For longer chains there was a drop in activity, typical for the substrate preference for an esterase. As a general rule, esterases have been defined

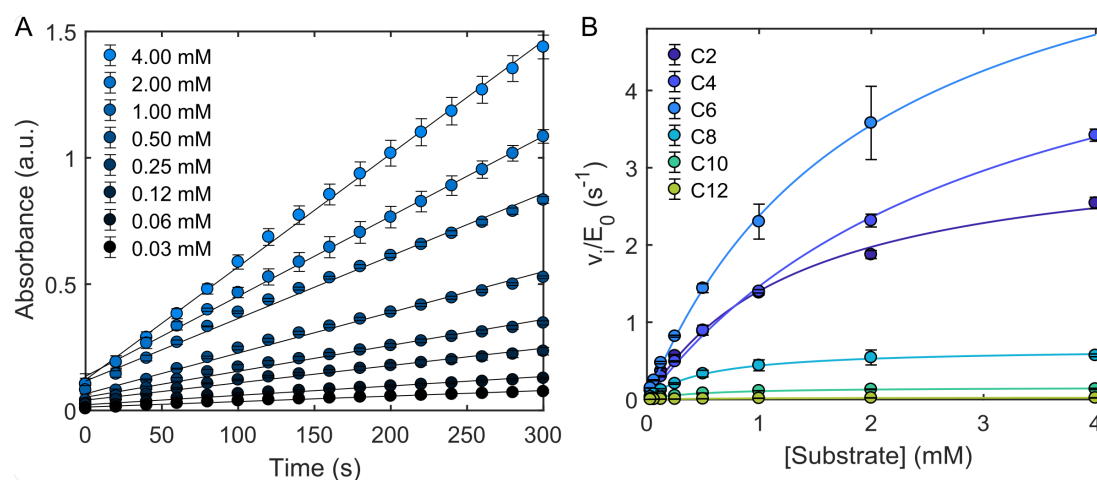


Fig. 3.17 Michaelis-Menten kinetics for hit N7. A: The initial rates were determined at different concentrations of p-nitrophenyl acetate by linear regression. $[E_0] = 250$ nM, 50 mM TrisHCl pH 8, 100 mM NaCl, 0.3% TritonX100. Reactions were corrected for the background rate in buffer. Error bars are the standard deviation of a triplicate measurement. B: Initial rates were plotted against substrate concentration and Equation 3.1 used to determine the catalytic parameters. The chain length of the p-nitrophenyl carboxylate was varied to determine the substrate scope of the enzyme.

by their ability to hydrolyse ester substrates of up to 10 carbons. Those which preferentially hydrolyse substrates with chains beyond 10 carbon atoms are considered lipases [157].

Similar to N7, all hits showed their highest activities between 4 and 8 carbon atoms, which agrees with the fact that fluorescein dihexanoate was used to isolate these enzymes in the droplet screening (Figure 3.18). The majority of esterases isolated from metagenomic libraries using the tributyrin assay, show their highest activity towards p-nitrophenyl acetate [40, 122, 126, 129]. Therefore, the droplet screen using fluorescein dihexanoate gave access to enzymes with a substrate preference shifted towards longer ester chains. Hit N2 had detectable activity only on substrates from two to six carbon atoms. All other enzymes were active up to chain lengths of ten carbon atoms and five of them had detectable activity up to the longest chain length. For p-nitrophenyl hexanoate, *i.e.* at the same chain length as in the droplet substrate, N20 had the highest activity at $(2 \pm 1) \times 10^5 \text{ M}^{-1}\text{s}^{-1}$ and N11 the lowest activity at $(200 \pm 20) \text{ M}^{-1}\text{s}^{-1}$. Therefore, enzymes with activities spanning three orders of magnitude were isolated. The average catalytic efficiency was in the range of 10^4 placing the hits into the activity realm of the enzymes active in the fatty acid metabolism [158].

In terms of k_{cat} , there was a general decrease with increasing chain length (Figure 3.19). In terms of K_m , nearly all enzymes preferred the short chain esters. Notably, N7 preferred longer to shorter chain esters and had its lowest K_m for p-nitrophenyl palmitate at $(360 \pm 30) \mu\text{M}$,

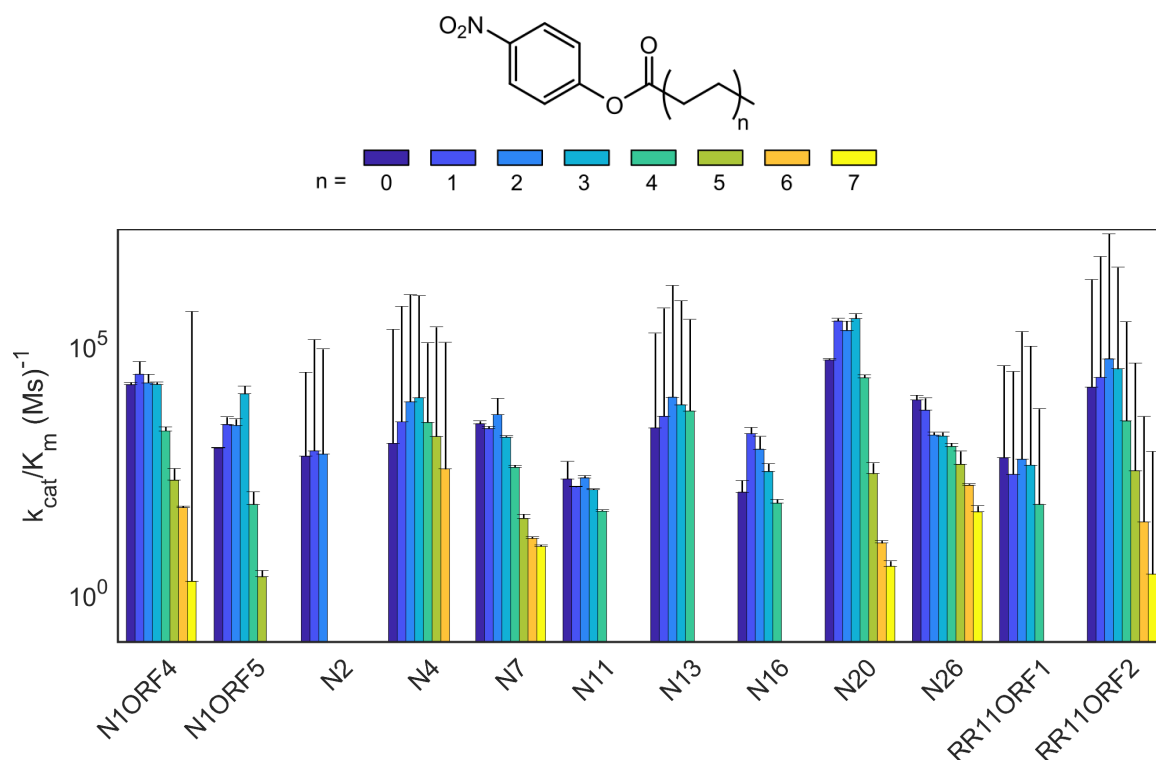


Fig. 3.18 Catalytic efficiencies of the metagenomic hits depending on the ester chain length. Error bars are the standard deviation from the Michaelis-Menten fit.

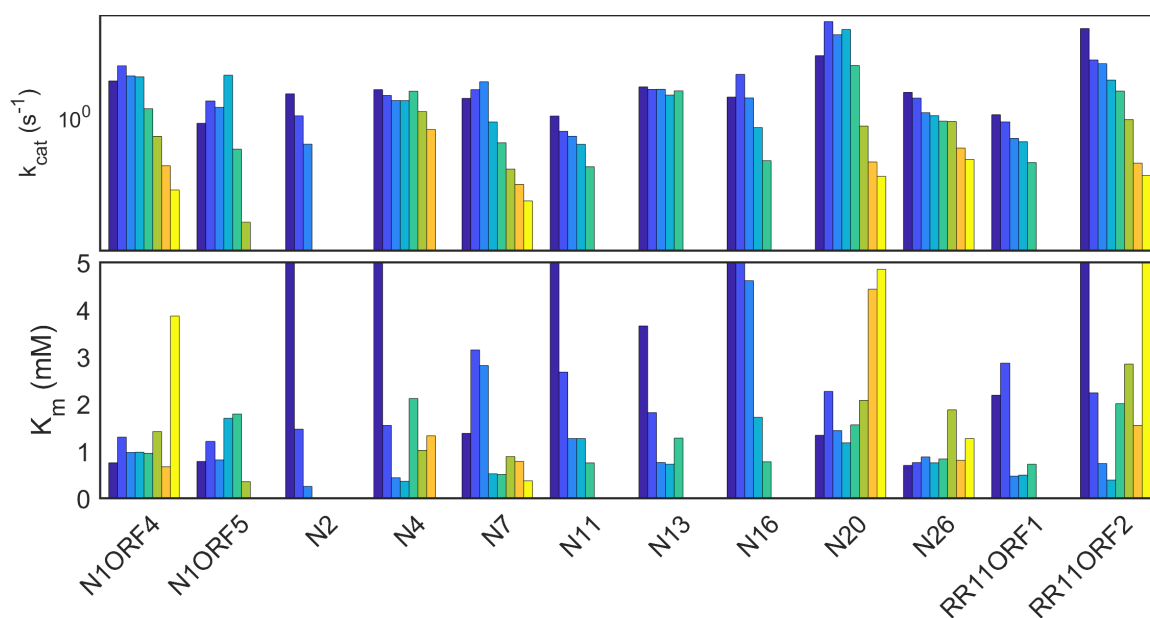


Fig. 3.19 Turnover number k_{cat} and Michaelis constant K_m of the metagenomic hits for p-nitrophenyl esters of different chain lengths (colour legend in Figure 3.18). Error bars omitted for clarity.

indicating that this enzyme has a large hydrophobic binding pocket. In summary, all the expressed metagenomic hits were shown to be esterases.

3.5.2 Melting temperatures

The melting temperature T_m of the hits was determined by thermal denaturation in the presence of the fluorescent dye SYPRO orange. The quantum yield of this dye increases upon exposure to a hydrophobic environment [159]. Therefore, as a protein unfolds and exposes its hydrophobic core, an increase in fluorescence is observed. The resulting melting curves are shown in Figure 3.20. A Boltzmann equation was fitted to the data assuming there were no intermediates between the folded and unfolded states [160]:

$$F(T) = \frac{F_{\text{end}} - F_{\text{start}}}{1 + \exp \frac{T - T_m}{C}} + F_{\text{start}} \quad (3.2)$$

With $F(T)$, F_{start} , and F_{end} being the signal depending on temperature T , at the start, and the end of the measurement respectively. The constant C characterises the breadth of the transition.

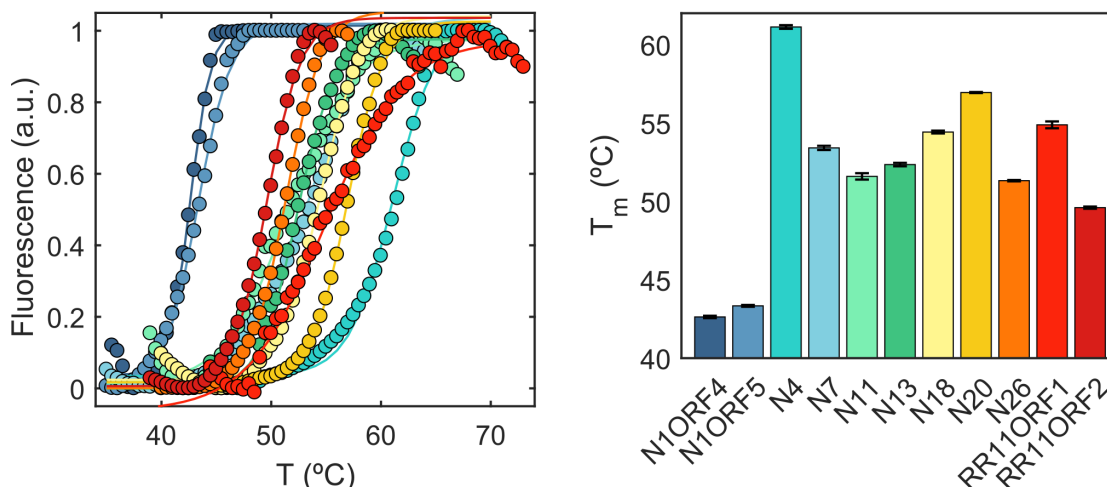


Fig. 3.20 *Left*: Shown are the melting curves from differential scanning fluorimetry of the metagenomic hits (average of three measurements). *Right*: A fit of equation 3.2 to the data yielded the melting temperatures T_m shown in the barplot. Errorbars are the standard deviation of the fit.

The obtained melting temperatures ranged from about 40 °C to 60 °C with an average T_m of (51 ± 5) °C across all the tested enzymes. These values are typical for mesophilic enzymes [161]. N1ORF4 and N1ORF5 have the lowest and very similar melting temperatures at

(42.63 ± 0.08) °C and (43.33 ± 0.06) °C. Both come from the same plasmid as their common origin, explaining the similar melting temperatures. The most thermostable enzyme was N4 at (61.1 ± 0.1) °C. Together, these melting temperatures are consistent with the screening conditions at room temperature.

3.5.3 Screening for promiscuous reactions

The catalytic promiscuity of the enzymes for different hydrolytic reactions was tested using a range of substrates with para-Nitrophenol as a leaving group. The hits N2 and N18 were omitted as they could not be obtained in sufficient quantities. Figure 3.21 shows all of the tested reactions. Column 1 to 8 tested for esterase (1), lipase (2,3), β -lactamase (4), amidase (5), phosphatase (6,7), and sulfatase (8) activities. Columns 9-14 tested for the hydrolysis of different glycosides. Column 17 tested for the Kemp elimination, and finally, columns 18 to 20 tested for thioesterase activity. With regards to the observed initial rate v_{init} during the first 3 min, a reaction was considered positive if it exceeded $0.3 \mu\text{M min}^{-1}$, which under these conditions corresponds to a change in absorbance of 0.01. Imperfect mixing at the start of the reaction can affect the value for the lower rates passing this threshold. Therefore, the concentration of product after 30 min was taken into account and the threshold set to $3 \mu\text{M}$. This corresponds to the concentration that should have accumulated after just 3 min given the speed threshold. An enzyme was only considered positive for a reaction if it met both these conditions.

As can be seen, in the bottom panel of Figure 3.21, all enzymes tested positive for the short-chain ester, as expected. All of them except for N1ORF5 and N11 also acted on the two long-chain esters. Furthermore, all hits except for N11 and N13 showed thioesterase activity. However, the hydrolysis of thioesters by esterases is not a surprising finding: thioesters are more reactive than the equivalent esters. This can be seen in the lower pK_{a} of thiols (*ca.* 11) than alcohols (*ca.* 16) making them better leaving groups. More interestingly than the above, the enzymes N16, N26, and RR11ORF2 showed β -lactamase activity (1), N1ORF5 showed β -galactosidase activity (2), and N7 showed Kemp eliminase activity (3).

In the literature, the ability of esterases to hydrolyse β -lactams is known but mostly discussed in the context of members of family VIII ([134]). These enzymes show homology with Class C β -lactamases such as AmpC of *E. coli* and are not members of the α/β -hydrolase family [133] but of the β -lactamase clan (Pfam CL0013). Some of them are only able to hydrolyse esters [162, 163], but some show activity towards both esters and β -lactams [164, 165]. However, neither of the three enzymes here are members of family VIII. N16 is a member of the carboxylesterase family (family VII), which also contains porcine liver esterase. Porcine liver esterase, a canonical esterase of great commercial importance, has already been shown to hy-

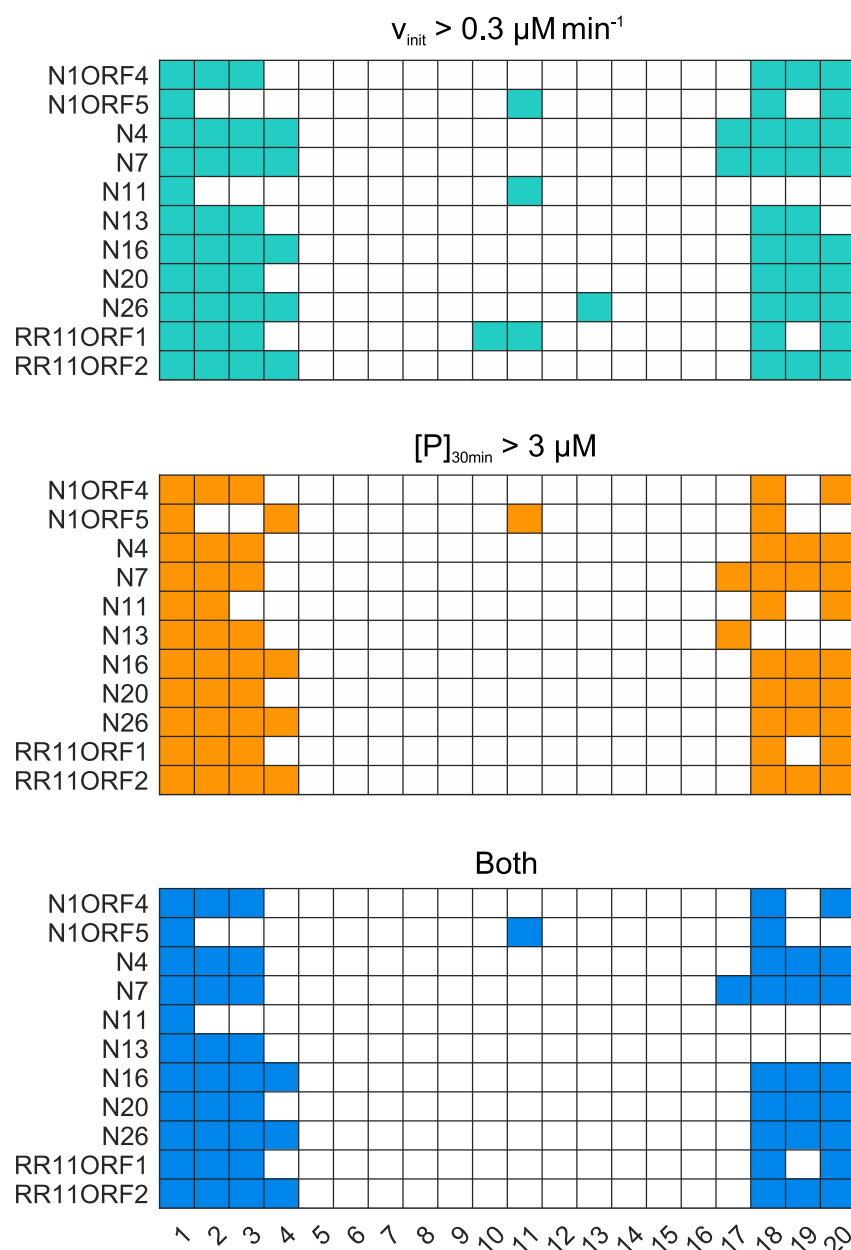


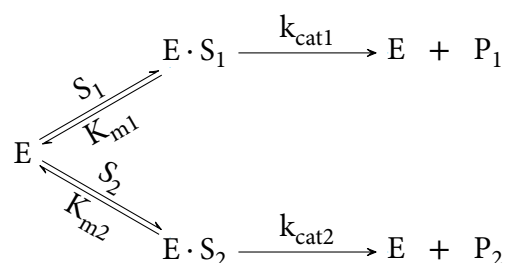
Fig. 3.21 Catalytic promiscuity test of the metagenomic hits. Coloured squares are reactions that exceeded the indicated thresholds for the observed initial reaction rate (*top*), for the amount of product that accumulated after 30 min (*middle*), and for both these conditions (*bottom*). All enzymes were used at the highest concentration possible and each substrate was at 1 mM. The substrates were: 1: pNP-hexanoate, 2: pNP-dodecanoate, 3: oNP-decanoate, 4: β -Lactam (CENTA[™]), 5: pNitroacetanilide, 6: pNP-Phosphate, 7: pNP-Phenylphosphonate, 8: pNP-Sulfate, 9: pNP- β -D-Glucopyranoside, 10: pNP- α -D-Glucopyranoside, 11: pNP- β -D-Galactopyranoside, 12: pNP- α -D-Galactopyranoside, 13: pNP- β -D-Xylopyranoside, 14: pNP- β -D-Fucopuranside, 15: 2-Chloro-4-phenyl- β -cellobioside, 16: pNP- β -Xylobiose, 17: 5-Nitro-1,2-benzisoxazole, 18: S-Phenylthioacetate, 19: 2,3-Mercapto-1-propanol, 20: S-Methyl-thiobutanoate.

drolyse a β -lactam ring selectively over an ester bond within the same molecule [166]. The families of N26 (C-C bond hydrolase) and RR11ORF2 (DUF3089) have not been reported to catalyse the hydrolysis of β -lactams.

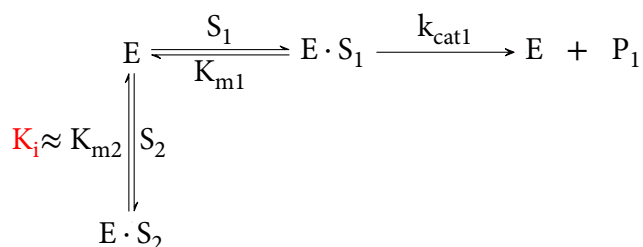
To our knowledge, the activities of esterases towards β -galactosides and Kemp substrates have not been reported previously. The Kemp Elimination is a reaction which will be discussed in detail in Chapter 4. Briefly, it is the general base catalysed abstraction of a proton from 1,2-benzisoxazole and its derivatives [94]. The reaction is deemed “*non natural*”, i.e. it is not known to play a role in the metabolism of any life form. Any protein found to catalyse the Kemp elimination is by definition promiscuous towards it. To our knowledge, no esterases have yet been reported to catalyse this reaction.

The catalytic parameters of β -galactosidase activity of N1O5 and the Kemp eliminase activity of N7 were determined. They were found to be $(3.0 \pm 0.8) \text{ M}^{-1}\text{s}^{-1}$ and $(5 \pm 1) \text{ M}^{-1}\text{s}^{-1}$ respectively. This is 100 to 1000 fold lower than their respective esterase activities and typical for a promiscuous activity.

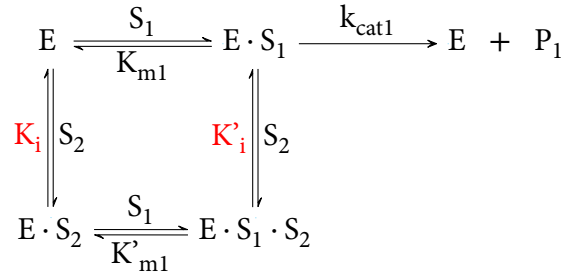
To investigate if the two reactions were catalysed by the same active site, cross-inhibition experiments were performed. Van Loo *et al.* argued that if the k_{cat} differs substantially between two reactions, the substrate of one acts as an inhibitor of the other [167]. Consider the reaction of enzyme E with the two competing substrates S_1 and S_2 :



If $k_{\text{cat}1} \gg k_{\text{cat}2}$, then S_2 acts as an inhibitor of the reaction of E with S_1 . For N1O5 $k_{\text{cat}1} > 1000 \cdot k_{\text{cat}2}$ and for N7 $k_{\text{cat}1} > 100 \cdot k_{\text{cat}2}$. A 100 fold difference was sufficient to observe cross-inhibition in the case of van Loo *et al.* [167]. If both reactions are catalysed at the same site, the above reaction scheme can be simplified to a competitive inhibition model:



With K_i being similar to the Michaelis constant K_{m2} . If S_1 and S_2 react at different sites, S_2 would bind to both the free enzyme E and to the enzyme-substrate complex $E \cdot S_1$ and mixed inhibition would be observed according to:



The Michaelis-Menten equation for mixed inhibition takes both cases into account:

$$v_i = \frac{v_{\max} [S_1]}{K_{m1} \left(1 + \frac{[S_2]}{K_i} \right) + [S_1] \left(1 + \frac{[S_2]}{K'_i} \right)} \quad (3.3)$$

Competitive inhibition is a special case of mixed inhibition where $K'_i \gg [S_2]$ and therefore $\frac{[S_2]}{K'_i} \approx 0$. Non-competitive inhibition is another special case where $K_i = K'_i$, *i.e.* binding of S_1 does not change the affinity of the enzyme towards S_2 and would indicate that the esterase and promiscuous reactions are independent.

To distinguish between these cases, the Michaelis-Menten kinetics of N1O5 and N7 were determined for S_1 at varying concentrations of their respective promiscuous substrate S_2 (Figure 3.22). Equation 3.3 was globally fitted to the data to determine K_i and K'_i .

For N1O5, the esterase reactions was measured in the presence of 0 to 10.5 mM pNP- β -D-galactopyranoside (S_2) with 0.3 to 6.3 mM pNP-hexanoate (S_1). The K_i was determined to be (2.1 ± 0.7) mM and K'_i was (18 ± 9) mM indicating mixed inhibition. This could be explained by the two substrates reacting at two different sub-sites of the active site or by binding of S_2 at a different site and inducing a conformational change in the enzyme. If S_1 and S_2 interact with different sub-sites in the active site, this could hint at a natural substrate that contains both a sugar and an ester moiety (*e.g.* an *o*-acylated sugar, which are abundant in plants [168]). Which of these two hypotheses is the case could be investigated by co-crystallisation experiments of N1O5 with β -galactose, para-nitrophenol and an esterase inhibitor.

In the case of N7, the esterase reaction was measured in the presence of 0 to 1.35 mM 5-nitro-1,2-benzisoxazole (S_2) with 0.3 to 6.3 mM pNP-butyrate (S_1). K_i was found to be

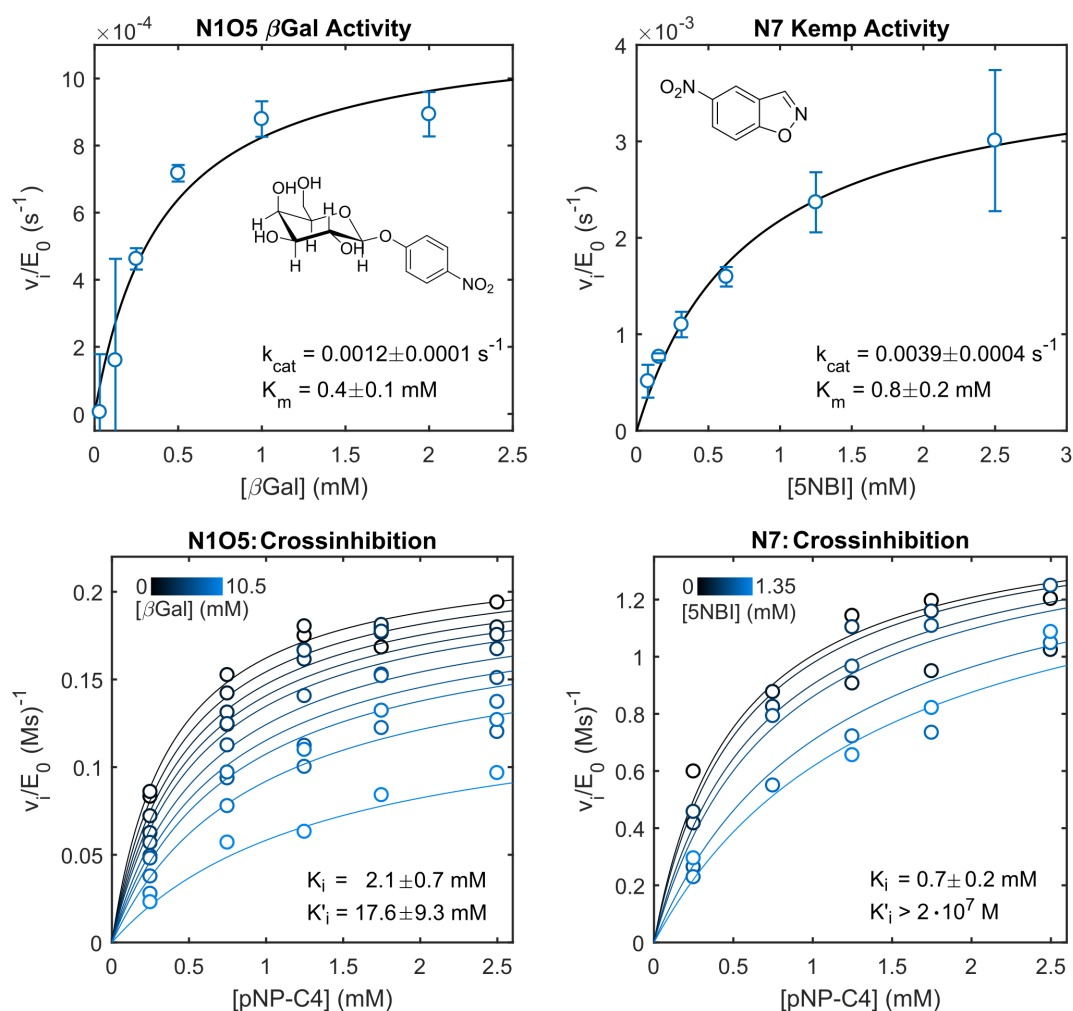


Fig. 3.22 The Michaelis-Menten parameters were determined for the promiscuous activities of N1O5 and N7 towards pNP- β -D-galactopyranoside and 5-nitro-1,2-benzisoxazole, respectively. Cross-inhibition experiments were performed in which the esterase reaction was measured in the presence of the promiscuous substrates and Equation 3.3 was fitted to the data to obtain the inhibition constants. The large K'_i for N7 indicates competitive inhibition between the two substrates (*i.e.* catalysis within the same active site environment), whereas N1O5 shows mixed inhibition (indicating catalysis at sites).

(0.7 ± 0.2) mM and $K'_i > 2 \times 10^7$ M. Because $K'_i \gg [S_2]$, the term $\frac{[S_2]}{K'_i} \approx 0$. This suggests that competitive inhibition of the esterase activity by 5-nitro-1,2-benzisoxazole took place ($K_i \approx K_{m2} = (0.8 \pm 0.2)$ mM), *i.e.* that the esterase and Kemp elimination reactions were indeed catalysed by the same enzyme active site.

The Kemp elimination is catalysed by general bases, whose reactivity can be significantly increased by non-specific and specific medium effects provided by the interior of enzyme active sites [169]. It is thus likely that the active site of N7 provides a hydrophobic environment in which a reactive general base exists. The nucleophile of the catalytic triad in N7 is not necessarily responsible for the general base catalysis. Ketosteroid isomerase has a comparable promiscuous activity of $2.5 \text{ M}^{-1}\text{s}^{-1}$ towards 5-nitro-1,2-benzisoxazole [170]. Mutation to alanine of the general base involved in the native reaction led to an increase of k_{cat}/K_m to $1.7 \times 10^5 \text{ M}^{-1}\text{s}^{-1}$. A nearby residue, not involved in the catalysis of the native reaction, was identified to catalyse the promiscuous reaction. Whether N7 is a similar case could be tested in future work by mutation of the residues in the catalytic triad. N7 also provides an interesting starting point for the evolution of a new non-natural function and may thus prove to be an interesting enzyme for directed evolution. While man-made catalysts of the Kemp elimination have been subjected to directed evolution, an environmental enzyme with natural promiscuity has not. The implications of such an experiment in the study of enzyme evolution are discussed in more detail in the introduction of the next chapter.

3.6 Conclusion

This chapter reports the first functional metagenomic screening for esterases using droplet microfluidics. In total, over 30 million droplets were sorted, amounting to $10\times$ coverage of a metagenomic library consisting of one million members. This throughput makes it the study with the highest number of clones screened for esterases to date. Furthermore, the droplet sorting itself consumed only about 100 μL of aqueous reagents, about 15 μg of surfactant, and less than 5 mL of fluorinated oil per sorting campaign. In terms of consumables, sample preparation at the macroscopic level thus exceeded the cost of droplet screening, making this technology very resource-efficient.

Thirteen esterase genes were discovered of which a large proportion was from small esterase families with few members characterised to date. All the identified genes were shown to encode for esterolytic enzymes. Their preference was for esters with chain lengths of 4 to 8, in agreement with the screening conditions.

In the field of directed evolution, Frances Arnold famously stated that “*you get what you screen for*” [171], which was the case in this screening campaign. Sometimes one also gets

something in addition to what was screened for. A test for promiscuous activities of the esterases showed that 3/11 enzymes exhibited β -lactamase activity, 1/11 showed β -galactosidase activity, and 1/11 showed Kemp eliminase activity. This adds to the evidence that enzymes are able to catalyse more than one reaction. It also shows that a substrate can be used as “bait”, to enrich for enzymes which have a certain desired catalytic property. Fluorescein dihexanoate can be seen as a bait for enzymes with a nucleophile, because a nucleophilic attack is required to release the fluorophore. Fluorescein, with its aromatic xanthene and benzene moieties, also probes active sites for a hydrophobic environment, *e.g.* by interacting with aromatic side chains, which may enrich for activities such as the Kemp eliminase activity of hit N7.

Taking this concept further, the accumulation of plastics in the environment has recently attracted the attention of the public. The first organism shown to degrade polyethylene terephthalate (PET) was isolated in 2016, indicating that microorganisms have started to evolve enzymes for the degradation of even recalcitrant plastics [172]. PET, and other polyesters such as poly-lactic acid, are difficult to screen directly at high-throughput because of their insolubility and the requirement to perform high-performance liquid chromatography (HPLC) to obtain a readout. However, pre-screening a large library for esterase activity should enable the enrichment for enzymes with such activities. Indeed, initial tests suggest that at least one of the identified hits has hydrolytic towards PET, thus providing a starting point for the directed evolution of improved activity.

The above quote also highlights the fact that in many enzymatic screening campaigns, the established assay can reduce the diversity of hits because of inadvertently co-selecting for properties other than the one of interest, for example the level of soluble protein expression.

In this functional screen, the diversity of the obtained hits may have been decreased by the need to perform a secondary screen on culture plates. The culture plate assay was required because more colonies were recovered than could have been re-screened in well plates. A high number of false positives was not surprising since it had been observed in the sorting campaigns for phosphotriesterase previously performed by Colin [84]. After one round of droplet screening, 4 hits were identified amongst 7,000 clones during the re-screening on plates, giving a hit rate of 6×10^{-4} . Here, the hit rate during re-screening for the esterases was 2×10^{-3} . The variation in hit-rate between experiments is not known, but the difference still gives an indication that esterases are more abundant in the SCV library as one would for an enzyme that is essential to every organism.

The absolute hit rate may have been decreased by the culture plate assay, because the clones needed to hydrolyse macroscopic amounts of tributyrin to be detected as positives. This makes the assay less sensitive compared to well plate assays where micromolar (ab-

sorbance) and nanomolar (fluorescence) concentrations of product are routinely detectable using optical methods. They also only provide a qualitative yes/no readout with a detection limit difficult to quantify because it depends on the possible incubation times, colony size, enzyme activity, amount of enzyme expressed, and judgement of the experimentalist. Some colonies without halos were tested in the cell lysate assay, but none of them were found to be active, confirming that the majority of colonies were “true” false positives. However, it can not be ruled out, that some hits were lost because of the lack of sensitivity.

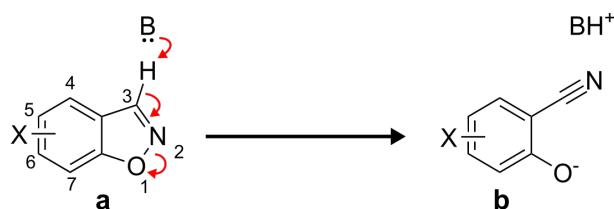
There are several sources of technical and biological bias, which can explain the high rate of false positives and will be discussed in more detail in final conclusions (Chapter 6), where observations made in Chapter 4 will be taken into account.

Chapter 4

An absorbance-activated droplet sorting assay for the Kemp elimination

4.1 Abstract

In this chapter, I present an absorbance-activated droplet sorting assay to screen for Kemp eliminases in droplets. The Kemp elimination is an important model reaction of proton transfer at carbon (Figure 4.1). It has been used extensively to build chemical and biological systems to mimic and study catalysis by enzymes. It is considered a non-natural reaction and only a few classes of natural enzymes are known to catalyse this reaction *via* promiscuous activities. So far, promiscuous enzymes catalysing the Kemp reaction have been found one-by-one. Finding them may have been due to intelligent searches (*e.g.* for hydrophobic binding pockets) or due to the fact that Kemp catalysts are widespread and therefore easy to find. The latter question can be addressed by employing ultrahigh-throughput functional metagenomic screening for Kemp catalysts in droplets.



1 X = H **2** X = 5-NO₂ **3** X = 6-NMe₂, 7-NO₂ **4** X = 6-SO₃H, 7-NO₂ **5** X = 5-NH₂ **6** X = 5-N₃

Fig. 4.1 The Kemp elimination is the general base catalysed elimination of a proton from 1,2-benzisoxazole. Number **1a** refers to the 1,2-benzisoxazole with X = H, **1b** refers to the respective Kemp reaction product.

The assay I established in this chapter was based on 5-nitro-1,2-benzisoxazole (**2a**), the most widely used substrate among substituted 1,2-benzisoxazoles. Its reaction product can be detected by virtue of its strong absorbance at 380 nm. The concentration of product was found to equalise in mixtures of droplets with different starting concentrations on a timescale of hours, *i.e.* the product leaked from droplets, which limited the incubation time for enzymatic assays.

It was possible to enrich a known Kemp eliminase, HG3.17, over a negative control, based on the activity detected in single-cell lysates. While the absorbance-based assay was not sensitive enough for functional metagenomic screening, it was employed to screen substitution and InDel libraries of Kemp eliminase HG3.17 (the functional metagenomic screen was performed using a novel fluorogenic substrate **6a** and is presented in the next chapter). The screening of HG3.17 libraries confirmed the ability of the developed method to enrich active enzyme variants. While no catalytically superior mutants were isolated, the substitution library yielded variants that were likely to have improved soluble expression. The InDel libraries revealed five regions tolerant to insertions and deletions in HG3.17, one of which showed 4.5 fold improved soluble expression. These near-neutral variants may open unprecedented mutational trajectories in future directed evolution studies.

Contributions: *All of the microfluidic work, the assay optimisation, the leakage study, the enrichment study, the construction of the libraries, screening, re-screening, hit identification, protein expression, purification and kinetic characterisation, data analysis, interpretation, presentation as well as structural interpretation of the identified mutations is entirely my own. Dr Josephin Holstein provided support by synthesising compounds **3a** and **4a**. Dr Tomasz Kaminski provided support with the final design of the in-line droplet sorter (the idea, initial designs and all of the experimental characterisation are mine).*

4.2 Introduction

In this introduction, I will first argue that identifying natural catalysts of the Kemp elimination has implications in the study of the evolution of enzymatic activities. I will then review the state of the research field focused on designing chemical and biological catalysts for the Kemp elimination and recent advances in finding promiscuous enzymes with high catalytic activities. I will end the introduction with the goals of this chapter and the anticipated challenges.

4.2.1 The Emergence of Phosphotriesterases

How enzymes with a new catalytic function arise is a central question in the study of molecular evolution. The earliest enzymes are thought to have been broad-range catalysts, which specialised *via* gene duplication and divergence [89, 173]. Present day enzymes remain able to catalyse more than one reaction and are thought to evolve new functions *via* such promiscuous activities [4, 6, 174].

The emergence of a new enzyme can be remarkably fast and was observed when synthetic chemicals were first introduced into the environment in the 20th century. One example is the insecticide parathion. It was developed in the 1940s by Schrader and was in widespread use from the 1950s until recently [175]. Parathion is converted to the phosphotriester paraoxon *in vivo* which inactivates acetylcholin-esterase. Both are types of compounds which do not occur naturally, but were soon observed to be degraded by microorganisms in soil and water. In the early 1970s strains of a *Flavobacterium* sp. and *Brevundimonas diminuta* were isolated that had the ability to hydrolyse the insecticide [176, 177]. Both encoded the same phosphotriesterase (PTE). Astoundingly, the enzyme catalysed the hydrolysis of paraoxon at the diffusion limit [178]. This *perfect* enzyme had evolved within the two decades after the first exposure of microbial communities to parathion. Other examples of enzymes and entire new metabolic pathways degrading anthropogenic compounds are known to have arisen during the same period [92].

Identifying the ancestral enzymes of these innovations is not trivial even in the case of such recent divergence [92]. In the case of the PTEs it was indeed a promiscuous lactonase activity, which led to the discovery of their progenitors: lactonases involved in quorum sensing [179]. The sequence identities between the two groups of enzymes are only at about 30% and it remains a challenge to reconstruct the evolutionary pathway that connects these lactonases to the PTEs [180].

4.2.2 How To Find Starting Points of Evolution

To better understand such evolutionary events, an interesting experiment would be to challenge a microbial environment with a new synthetic compound to detect and isolate enzymes which are potential starting points for the acquisition of new function. One way of doing this in the laboratory is to screen a metagenomic library against such a compound. The principle was shown by Colin *et al.* in an ultrahigh-throughput droplet screen for PTEs [67]. Eight clones encoding nine new PTEs were found in the metagenomic library SCV which consists of more than one million members (see Table B.1 for its composition). Their ability to hydrolyse phosphotriesters was likely to be a promiscuous rather than their native activity [67].

They may serve as starting points for further evolution and improvement of that activity. However, because organophosphates were and are used globally as insecticides, it cannot be ruled out that these activities had already evolved in these enzymes before they were isolated.

Applying this workflow to a synthetic substrate that microorganisms were not pre-exposed to would yield *ab initio* starting points of evolution.

4.2.3 The Kemp Elimination

A well studied model reaction is suitable for this task: the Kemp elimination [94, 95]. This is the general base catalysed proton elimination from 1,2-benzisoxazole yielding 2-hydroxy-benzonitrile, see Figure 4.2. The reaction proceeds readily in dilute aqueous solutions of the compound and several 5- or 6-substituted derivatives. Kemp and co-workers were interested in 1,2-benzisoxazol-3-ide as a leaving group in linear free energy relationships which provided access to a large range of reactivity without secondary effects due to structural changes [94, 95]. The reaction is susceptible to specific catalysis by hydroxide and general base catalysis but not acid catalysis [94]. The general base catalysis by acetate was found to be enhanced up to 10^7 fold in polar aprotic solvents compared to water, which can be ascribed to the desolvation of the carboxylate ion (Figure 4.2). Kemp *et al.* also argued that this effect would be key in the construction of enzyme-like catalysts for this reaction [181].

Indeed, it has been a long desired goal in enzymology to construct enzymes from first principles to escape the maxim of “*What I cannot create, I do not understand.*” (as left on Richard Feynman’s blackboard at the time of his death). The Kemp elimination is one reaction frequently used in attempts to achieve this goal. It is considered a model reaction for proton transfer from carbon, an often rate limiting step in enzymatic reactions [156]. Furthermore, there is no natural enzyme known to catalyse this reaction. The derivative **2a** was employed in most studies as it provides a convenient colorimetric readout ($\epsilon_{380\text{nm}}$ of $15.8 \text{ mM}^{-1} \text{ cm}^{-1}$ of the anionic product) and is activated compared to the unsubstituted 1,2-benzisoxazole (ΔH° of -39 and -35 kcal mol^{-1} respectively), thus allowing the observation of even small rate enhancements [94]. The reaction rates given in this introduction refer to **2a** unless otherwise stated.

The systems developed or discovered to catalyse the Kemp elimination can be broadly grouped into chemical systems, catalytic antibodies, engineered proteins, the *de novo* designed enzymes, and promiscuous enzymes. As I will discuss in the following, these were rationally designed and further improved by directed evolution. I will compare them to promiscuous catalysts, which promote the reaction by chance.

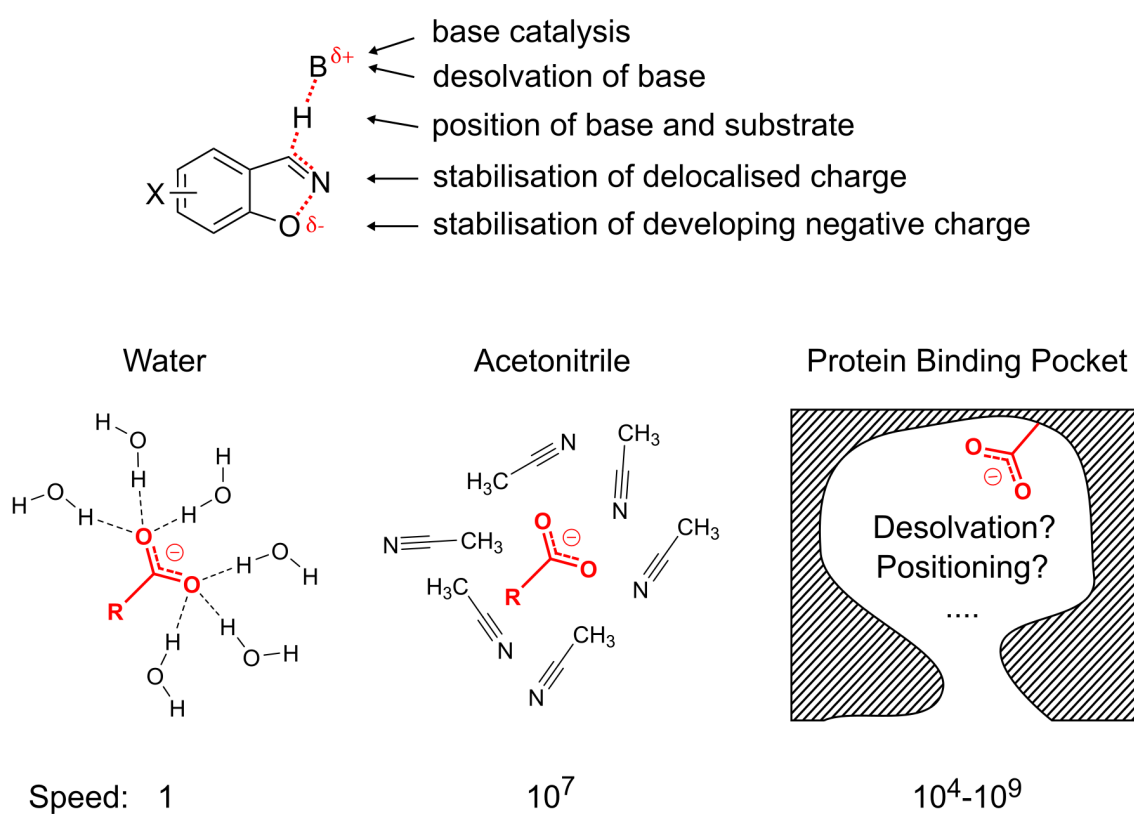


Fig. 4.2 Factors contributing to the catalysis of the Kemp reaction. If carboxylate is the general base, its reactivity is strongly enhanced in aprotic solvents. This effect can be exploited by desolvation of a carboxylate side-chain in the apolar environment of a protein binding pocket.

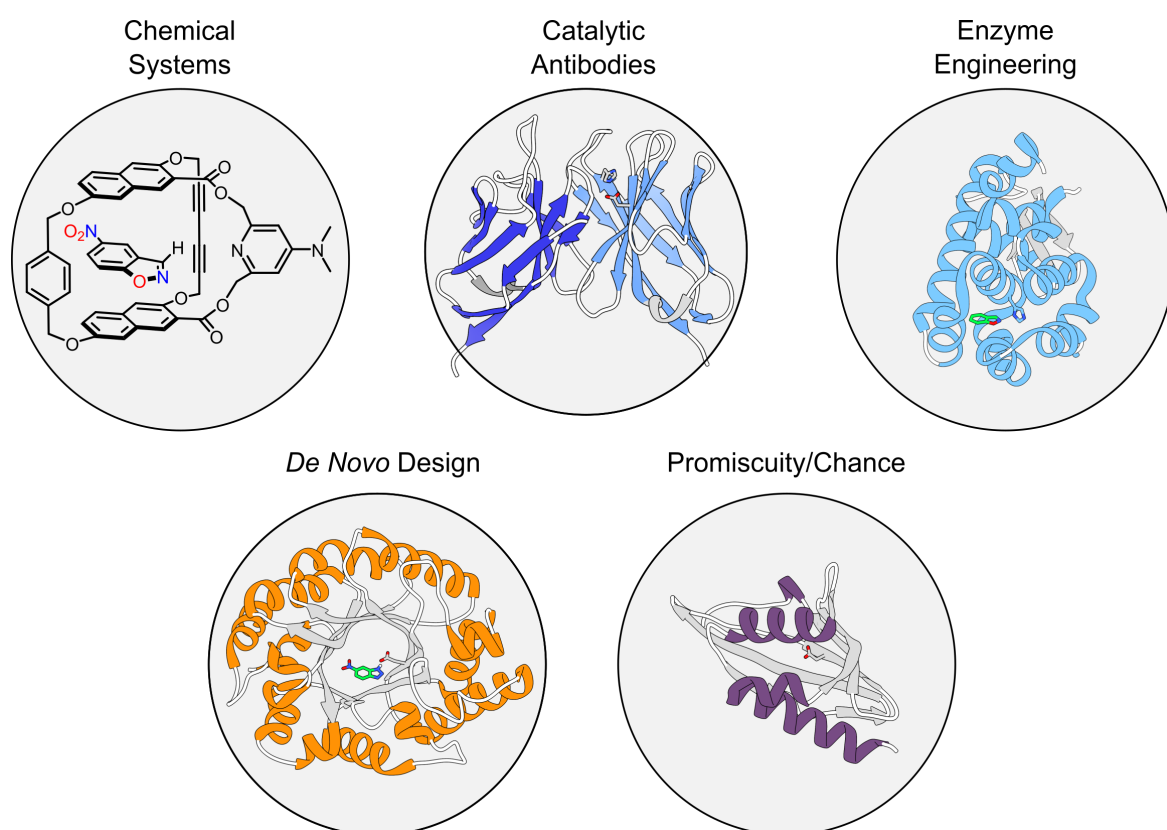


Fig. 4.3 Catalytic systems shown to promote this reaction fall into the five groups shown: chemical systems, catalytic antibodies, engineered enzymes, designed (and subsequently evolved) enzymes, and promiscuous enzymes.

Chemical Systems

Chemical host-guest systems bring the substrate into proximity with a catalytic group reminiscent of enzyme-substrate binding in the active site. One previously studied host is β -cyclodextrin (β CD), which was modified to yield amino- β CDs. A rate acceleration of up to 10^4 was reported if compared to the background reaction in buffer. However, if compared to the reaction of the free amine in buffer the rate was accelerated by only 10^2 fold [182]. The background rate which the observed reaction rate is compared to is inconsistent in the literature and has been a matter of debate. In the following, I will give rate accelerations compared to the free base, but this is not always a sufficient comparison as will be seen in the case of the catalytic antibodies.

The host-guest complex shown in Figure 4.3 was more successful than the β CDs: the rate acceleration of the host complex was over 10^3 compared to the free amine indicating that host binding contributed strongly to the observed catalytic effect [183].

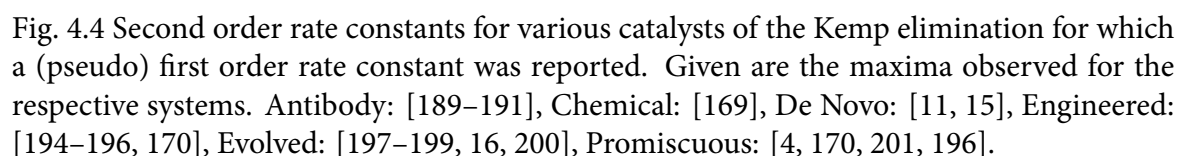
Another example is a coordinate cage which self-assembles in aqueous solution and accelerates the reaction, notably of unsubstituted 1,2-benzisoxazole, 10^5 fold. It binds the substrate in its cavity and recruits hydroxide ions to the cage surface as shown by the pH independence of the catalysed reaction [184].

This mechanism is similar to an even simpler system in which cationic surfactant-micelles catalyse the Kemp elimination [185]. This system was recently expanded by adding surfactant molecules with a phosphate or carboxylate head group which can act as a general base; rate accelerations of 10^4 were observed in both cases [186].

Finally, a chemical system based on methylated polyethyleneimine (synzymes) was synthesized by Hollfelder *et al.* [187, 169]. This enzyme mimic allowed analysis according to the Michaelis-Menten model and a rate acceleration of 10^6 was achieved. Notably, the catalytic amine groups were 100 to a 1,000-fold more reactive than in bulk solvent (either water or acetonitrile). This suggested that specific medium effects in the active site enhanced the reactivity of catalytic groups without the need for precise positioning relative to the substrate.

Catalytic Antibodies

Catalytic antibodies were the most advanced protein-based enzyme mimics generated until the 1990s. They are generated by raising an antibody against a transition-state mimic of the desired reaction [188]. Among the many reactions for which antibodies were raised was also the Kemp Elimination [189–191]. One of the most successful catalytic antibodies was 34E4 with a $k_{\text{cat}}/K_{\text{m}}$ of $5.5 \times 10^3 \text{ M}^{-1}\text{s}^{-1}$ [189]. Its catalytic residue was a glutamate and the rate acceleration compared to acetate in buffer was 3.8×10^8 [189, 192]. The authors explained



Enzyme Engineering

Two notable studies introduced Kemp eliminase activity into an existing protein scaffold using a rational approach. Merski *et al.* introduced Kemp eliminase activity into T4 lysozyme by creating a hydrophobic binding pocket employed with a histidine acting as the general base [194]. Korendovych *et al.* created “AlleyCat”, an allosterically regulated Kemp elimi-

nase, by means of a single point mutation, phenylalanine to glutamate, in calmodulin [195]. The second order rate constants were 1.8 and $5.8 \text{ M}^{-1}\text{s}^{-1}$ respectively.

***De novo* Enzymes**

The most advanced attempt at creating a catalytic activity from first principles in an existing protein scaffold was the design of the so-called *de novo* enzymes. In this approach, chosen catalytic residues were positioned around the transition state of the reaction and the geometry optimised to maximise transition state stabilisation [11]. The idealised active sites were then matched with protein scaffolds that could accommodate them. Röthlisberger *et al.* chose to express 59 of the resulting protein sequences, the KE series of Kemp eliminases, of which 8 had detectable activity. The second order rate constants ranged from 6 to $160 \text{ M}^{-1}\text{s}^{-1}$. Privett *et al.* followed a similar design strategy, but iteratively improved on their first design by incorporating experimental insights from kinetics and x-ray crystallography [15]. After two rounds of experimental feedback, they obtained HG3 with a second order rate constant of $430 \text{ M}^{-1}\text{s}^{-1}$.

The *de novo* catalysts are superior to the simpler engineering approaches, but they fall short of both the synzymes, which made no assumptions about the transition state geometry, and the catalytic antibodies, which were based on randomised libraries. The authors in both papers acknowledged this shortcoming and subjected their design outcomes to directed evolution to improve their enzymes.

The Success of Enzyme Design

Before commenting on directed evolution, it seems appropriate to assess the success of the four approaches discussed so far. All aimed to bring about catalytic activity based on the known principles of enzyme catalysis and detailed knowledge of the Kemp elimination. Most of the Kemp eliminases discussed use a carboxylate as the catalytic base. The success of these designed enzymes can thus be benchmarked against the second order rate constant k_2 of carboxylate in water and acetonitrile, respectively (10^{-4} and $10^3 \text{ M}^{-1}\text{s}^{-1}$, respectively) [193, 202]. As is evident from Figure 4.4, all of the designed enzymes have a second order rate constant much improved over carboxylate in buffer (10^4 - 10^8 fold). However, independent of the detailed factors contributing to these enhancements, none of them exceed what is achieved by simply desolvating carboxylate.

Directed evolution produced the most powerful, experimentally-derived Kemp eliminases to date [197–200, 16]. Two KE variants and HG3.17 markedly exceed the benchmark rate of carboxylate in an organic solvent. HG3.17 holds the current record at $2.3 \times 10^5 \text{ M}^{-1}\text{s}^{-1}$

after an impressive 17 rounds of directed evolution alternately using random mutagenesis and recombination of improved variants. In the literature, these final enzymes are often referred to as computationally-designed, but in truth their current second order rate constants are thanks to the power of directed evolution. The use of directed evolution changes the central problem from a protein design challenge to one of choice: the choice of the starting point and the choice of method for diversification and screening. In fact, the five Kemp eliminases that have been evolved reach final second order rate constants in strict order of what the respective starting efficiency was with a typical increase of 2-3 orders of magnitude. There is no obvious reason of why the starting point needs to be a *de novo* designed protein. The starting point could be a natural, promiscuous enzyme.

Promiscuous Enzymes

The serum albumins were the first proteins investigated systematically for their promiscuous Kemp eliminase activity [193, 203, 204]. The highest reported second order rate constant for BSA was $6.5 \times 10^3 \text{ M}^{-1}\text{s}^{-1}$ at pH 10 and above – which puts it on a par with catalytic antibody 34E4. Khersonsky *et al.* performed the first systematic screening for Kemp eliminase activity [205]. Screening the ASKA library¹, two promiscuous enzymes were identified: YdbC, an oxidoreductase ($430 \text{ M}^{-1}\text{s}^{-1}$), and XapA, a phosphorylase ($10^3 \text{ M}^{-1}\text{s}^{-1}$). In 2017, three groups reported promiscuous Kemp enzymes which were chosen to be studied based on rational arguments. Lamba *et al.* selected ketosteroid isomerase (KSI) because it is known to catalyse a proton transfer using a carboxylate of a hydrophobic substrate [170]. Indeed, KSI was found to have promiscuous activity of $2.5 \text{ M}^{-1}\text{s}^{-1}$. Remarkably, the authors mutated the catalytic aspartate of the ketosteroid reaction to asparagine and found the promiscuous activity was subsequently $1.7 \times 10^4 \text{ M}^{-1}\text{s}^{-1}$, *i.e.* by mere chance they created one of the most active Kemp eliminases known. Docking simulations suggested that a different aspartate in the active site is catalysing the Kemp elimination in KSI [170].

Notably, two studies reported promiscuous Kemp elimination in enzymes which employ heme as a cofactor. In the first study, Oxd was investigated [201]. These enzymes coordinate the substrate using the heme cofactor, see Figure 4.5A. The native nitrile product is obtained *via* N-O bond cleavage followed by proton abstraction. As the authors highlight, this is reminiscent of the Kemp elimination. Therefore, they tested three Oxds and all were active towards **2a** with second order rate constants of $10^2 \text{ M}^{-1}\text{s}^{-1}$ with iron(III) and 10^4 to $10^5 \text{ M}^{-1}\text{s}^{-1}$ with iron(II). Mutation of either the general base or the iron-coordinating histidine abolished catalytic activity. The authors concluded that both iron in its active, *i.e.*

¹The ASKA library is composed of variants that allow overexpression of every predicted ORF present in the *E. coli* genome [206].

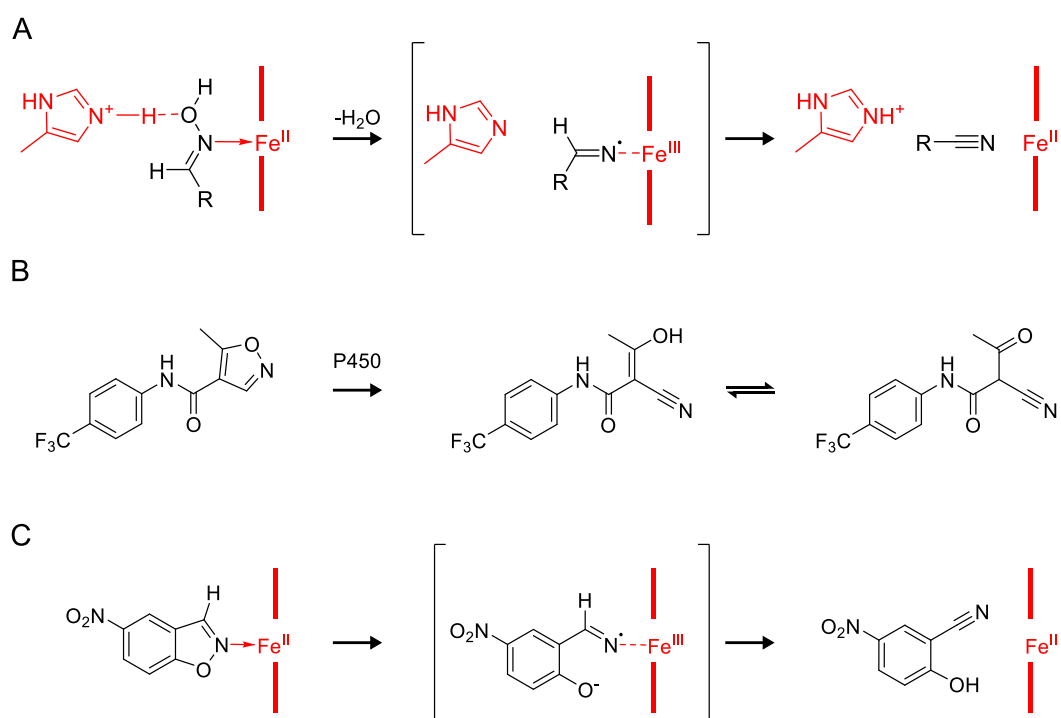


Fig. 4.5 Recently, aldoxime dehydratase (Oxd) and cytochrome P450 monooxygenase were shown to catalyse the Kemp elimination [201, 196]. *A*: In the proposed mechanism of Oxd N-O bond cleavage occurs before abstraction of the proton by histidine. *B*: The degradation of the drug leflunomide by P450 monooxygenase is formally equivalent to the Kemp elimination. *C*: The suggested mechanism of P450 monooxygenase involves N-O bond cleavage before elimination of the proton takes place.

reduced, state and the general base were required for efficient catalysis. However, they did not further study the mechanism of the promiscuous activity [201]. Li *et al.* reported another heme-containing enzyme to promiscuously catalyse the Kemp reaction: cytochrome P450 monooxygenase [196]. This enzyme catalyses the isoxazole ring scission in the drug leflunomide (Figure 4.5B), which is – again – formally equivalent to the Kemp elimination. Therefore, they tested P450-BM3 and found it catalysed the Kemp elimination with a second order rate constant of $240 \text{ M}^{-1} \text{ s}^{-1}$. A mutation known to enhance the activity of P450 with small molecule substrates boosted the second order rate constant to $3 \times 10^4 \text{ M}^{-1} \text{ s}^{-1}$. The authors suggest the redox mechanism in Figure 4.5C and provide experimental and computational evidence. Notably, there is no general base near the iron centre. Also, when iron(II) was replaced with zinc(II) there was no catalytic activity and addition of CO reduced the activity 10 fold, indicating that the reaction takes place at the iron centre.

In summary, a number of promiscuous enzymes have been found that rival the activity of designed and experimentally evolved Kemp eliminases. Most remarkable is the high activity without mutations of the three Oxds. The use of heme as a cofactor was never even considered in the design approaches. It is conceivable that in a large scale screening for promiscuous Kemp catalysts innovations could be identified that are not yet within reach of the decisive *chemical intuition* that guided the studies of Li and Mao.

4.3 Overview of this Chapter

In light of the importance of the Kemp elimination in the study of enzyme catalysis and design, it was meaningful to develop a droplet-based Kemp assay to enable higher throughput in library screening campaigns. The recently developed AADS allowed the implementation of the colorimetric Kemp assay with **2a** (see Figure 4.6). [61] This substrate yields the strongest reported absorbance upon turnover and is thus the most widely used in screening assays ($\epsilon = 15.8 \text{ mM}^{-1} \text{ cm}^{-1}$ at 380 nm).

In the previous chapter fluorescein was used as the optical reporter. While there had been previous studies showing the retention of fluorescein within microfluidic droplets enabling library screening [69, 67], it was not known if this would be the case for 5-nitro-2-hydroxy-benzonitrile. A number of challenges were anticipated and addressed (see also Figure 4.7:

1. The possible incubation time is limited by the spontaneous turnover of the substrate ($k_{\text{OH}^-} = 15 \text{ M}^{-1} \text{ s}^{-1}$, [94]) and general base catalysis of the buffer. For example, assuming a k_{buffer} of $10^{-4} \text{ M}^{-1} \text{ s}^{-1}$ and a typical concentration of 50 mM at pH 7, the half life of the reaction would be 27 h. This would be reduced to 17 h by using the buffer at

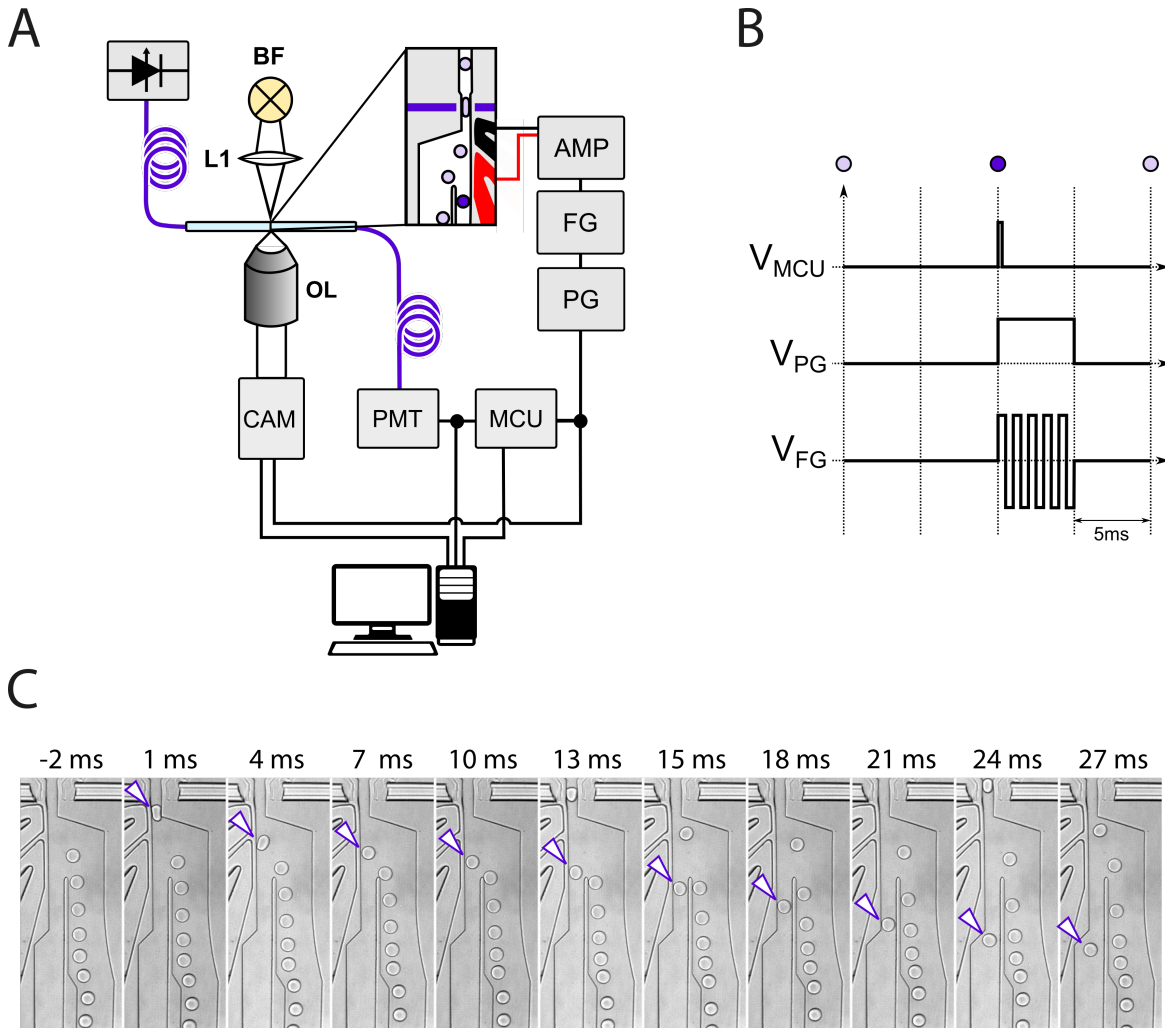


Fig. 4.6 Principle of AADS, which was based on [61]. *A*: Optical fibres (purple) were inserted into the sorting chip to measure absorbance across a $50\ \mu\text{m}$ gap through which the droplets passed one by one. A light-emitting diode (385 nm) was coupled to one fibre, and a PMT for signal detection to the other. The sorting process was controlled by a microcontroller (MCU) and monitored using a custom LabView program. *B*: The pulse sequence used to sort droplets follows the same principle as in FADS (Figure 2.2). It was started by the (MCU), which caused the pulse generator (PG) to open a gate for the function generator (FG). The square-wave was amplified (AMP) and resulted in the sorting of a droplet by DEP. *C*: Individual frames of a video recorded by the camera for one sorting event. Droplets entered from the top and exited *via* the waste channel on the right. The droplet which triggered the sorting event was pulled to the collection channel on the left as indicated by the arrows.

100 mM. Therefore, the buffer conditions were optimised to guarantee the success of a screening campaign.

2. Small molecules have been found to exchange between droplets [70, 72]. This is particularly detrimental for the product of a reaction (product leakage): if it is lost to surrounding droplet compartments, the signal is reduced and positives may never emerge from the droplet population. The degree of molecular transport of the reaction product was quantified and minimised.
3. The optimisation of assay conditions was not sufficient to enable detection of Kemp eliminase activity in droplets. Therefore, the substrate was chemically modified in order to prevent leakage of the product. While exploring different substrates, a novel fluorogenic Kemp eliminase substrate was discovered, which did not leak and was used for metagenomic screening using FADS described in the next chapter.
4. The AADS assay was used to screen substitution and InDel libraries of Kemp eliminase HG3.17 using the substrate this enzyme had been evolved with before and screening much larger libraries than previously possible.

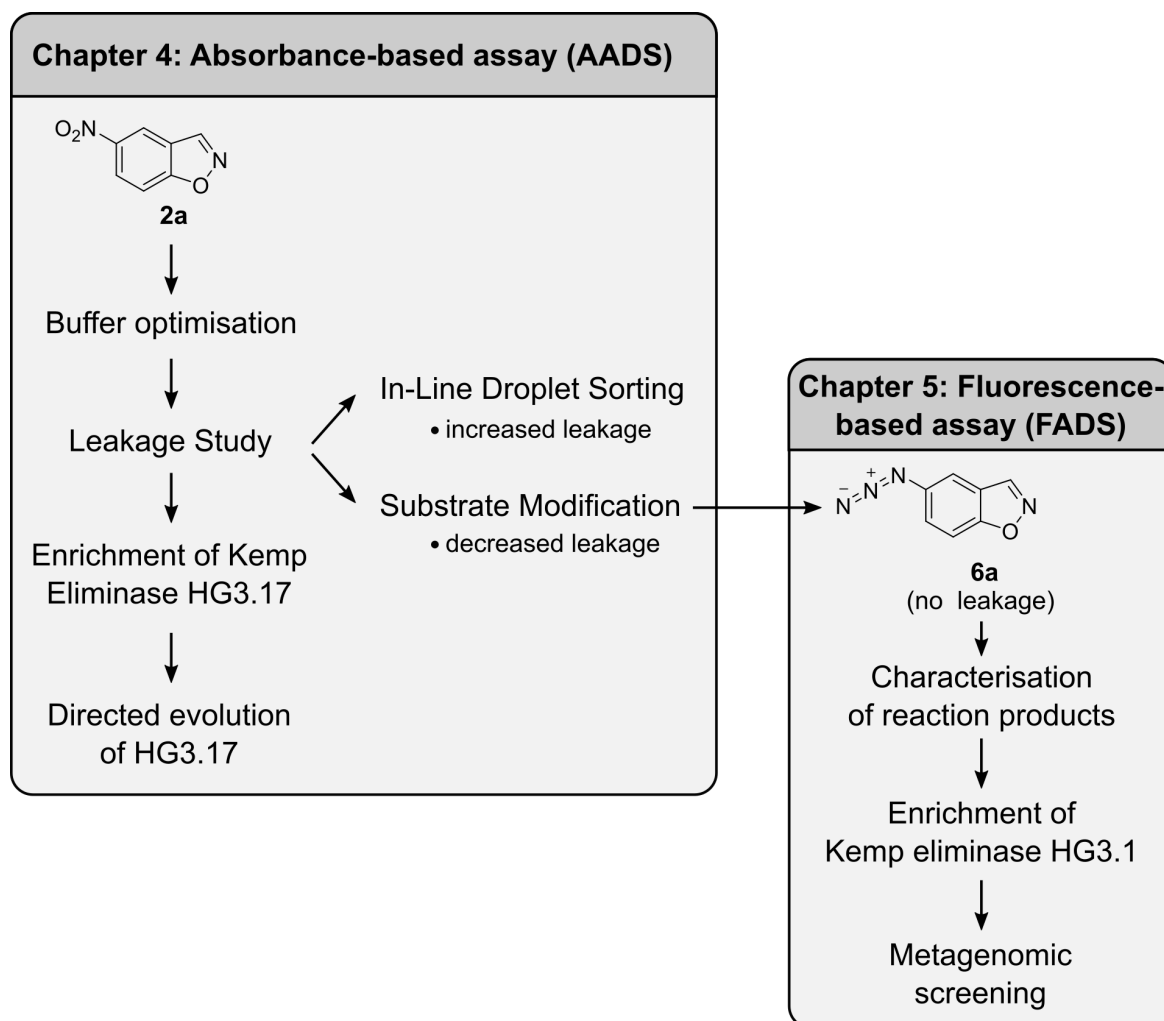


Fig. 4.7 Overview of the results section. First, substrate **2a** was investigated for its suitability to detect enzymatic activity in droplets. Its reaction product was found to transfer between droplets, *i.e.* to leak. Therefore, two strategies were followed with the aim to overcome this limitation: design of an in-line droplet generation and sorting device and modification of the substrate. One of the modified substrates, **6a**, was found to be fluorogenic and allowed the establishment of a fluorescence-based assay.

4.4 The absorbance-based Kemp eliminase assay

4.4.1 Establishment of buffer conditions and enzyme controls

The substrate **2a** was synthesised according to published procedures ([94]) and assay conditions were evaluated for an absorbance-based droplet screen. As explained above, the background reaction needed to be minimised to allow for long incubation times. Therefore, the general base catalysis of seven buffers was assessed, see Figure 4.8 and Table 4.1. Buffers are generally used near their pK_a . Using the equilibrium $HA + H_2O \rightleftharpoons H_3O^+ + A^-$ the pK_a is usually defined as:

$$pK_a = -\log_{10}(K_a) = -\log_{10}\left(\frac{[A^-][H_3O^+]}{[HA]}\right) \quad (4.1)$$

Assuming the concentration of H_2O remains constant. With $-\log_{10}([H_3O^+]) = pH$, this can be re-arranged to give the ratio of general base to general acid:

$$\frac{[A^-]}{[HA]} = 10^{pH-pK_a} \quad (4.2)$$

If the pH equals the pK_a , half the buffer molecules are protonated and half are not, which is where its ability to keep the pH constant is maximal. However, the general base of the buffering agent participates in the hydrolysis of 1,2-benzisoxazoles according to:

$$\frac{v_0}{[S]_0} = \underbrace{k_{H_2O}[H_2O] + k_{OH^-}[OH^-] + k_{A^-}[A^-]}_{k_{const}} \quad (4.3)$$

In a solution of the buffering agent in water, the initial rate of reaction v_0 divided by the initial substrate concentration $[S]_0$ is determined by the reactions with water, hydroxide and the general base. For the Kemp reaction it is thus useful to use the buffering agent below its pK_a . This has the advantage that the general base concentration is lowered and that the buffering capacity against hydroxide is enhanced. In 70 μm droplets, the cytosol of a single *E. coli* cell is diluted 20,000 fold and the surfactant used for cell lysis is mildly basic. Thus, the major downwards influence on pH is acidification caused by substrate turnover (**2a** has a pK_a of 4.1, [94]). As long as the general base is used in excess of substrate, this strategy should not be detrimental to the assay. At fixed pH, the first two terms in Equation 4.3 are constant (k_{const}) and k_{A^-} can be determined by varying the buffer concentration. The variable term of the observed reaction rate will be $k_{obs}[A_{tot}] = k_{A^-}[A^-] + k_{HA}[HA]$. Assuming k_{HA} is negligible at the given pH: $k_{A^-} = k_{obs}/([A^-]/[A_{tot}])$. The ratio $[A^-]/[A_{tot}]$ can be obtained

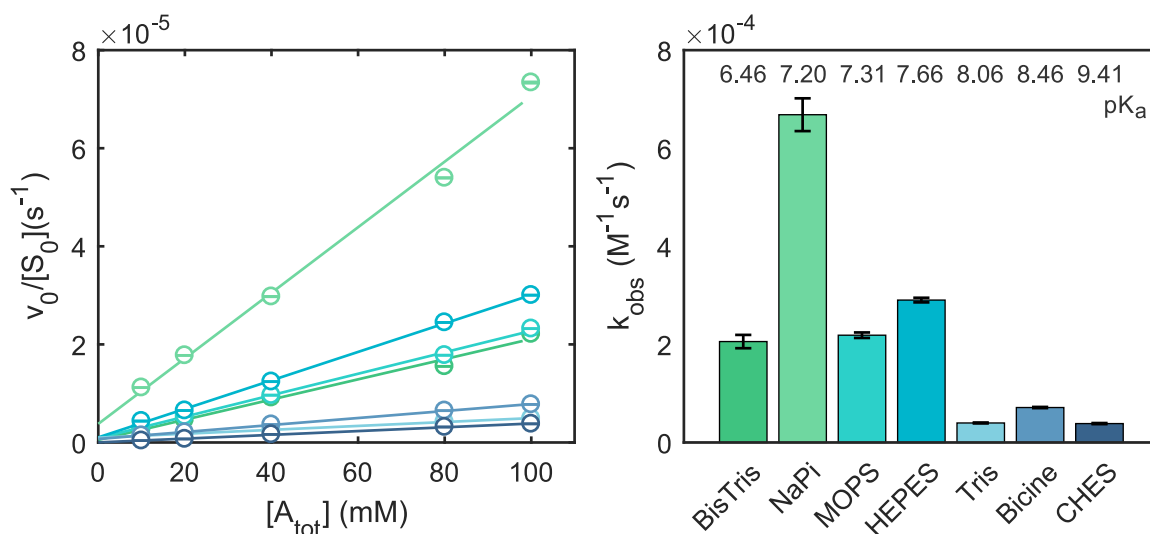


Fig. 4.8 Reaction rate of the Kemp elimination for **2a** in seven different buffers at pH 7. *Left:* The rate of reaction was dependent on the buffer concentration. Lines are linear regressions, error bars are the standard error of the slope of $v_0/[S_0]$. *Right:* Slopes (corresponding to k_{obs}) of the linear regressions. Sodium-phosphate buffer (NaPi) is most often used in studies of the Kemp reaction. However, of the tested buffers it caused the highest background reaction rate, even compared to BisTris, which has a lower pK_a . Error bars are the standard error of the slope estimate. $[S_0]$ was 1 mM and the reaction was followed at 380 nm.

from Equation 4.2 with $[A_{\text{tot}}] = [\text{HA}] + [\text{A}^-]$:

$$\frac{[\text{A}^-]}{[A_{\text{tot}}]} = \frac{10^{\text{pH}-\text{pK}_a}}{1 + 10^{\text{pH}-\text{pK}_a}} \quad (4.4)$$

According to this analysis, seven buffers were tested at pH 7 as shown in Figure 4.8. As expected, the initial rates were linearly dependent on the total buffer concentration. The mean of the intercepts yielded a k_{const} of $(1 \pm 1) \times 10^{-6} \text{ M}^{-1} \text{ s}^{-1}$, which is the minimum background rate possible and is in good agreement with the literature $2.1 \times 10^{-6} \text{ s}^{-1}$ for pH 7, calculated from [95]). The observed reaction rate constants were derived by linear regression and are given in the bar chart. Notably, the rate constant of the widely used NaPi buffer was 17 times higher than the one of Tris buffer. Table 4.1 shows the calculated k_{A^-} . BisTris and Tris have the lowest and second lowest rate constant in terms k_{A^-} . Note, however, that it is the observed reaction rate which determines the background rate in the final assay.

Next, HG3.17 (the Kemp eliminase with the highest reported catalytic efficiency to date [16], Appendix C.1 and C.3) was obtained as a positive control and cloned into plasmid pET32 under T7 expression. Human acid phosphatase 1 (ACP) in pET32 served as a negative control (plasmid and negative control courtesy of Dr Stephane Emond, Appendix Figures C.1

Table 4.1 Listed are the seven buffers tested for their reaction rate with the Kemp substrate **2a** and the measured reaction rate constants.

Buffer	pK _a (at 25 °C)	A ⁻ /A _{tot} (at pH 7)	k_{obs} (10 ⁻⁵ M ⁻¹ s ⁻¹)	k_{A^-} (10 ⁻⁴ M ⁻¹ s ⁻¹)
BisTris	6.46	0.776	21 ± 2	2.6 ± 0.2
NaPi	7.20	0.387	69 ± 7	17.3 ± 0.9
MOPS	7.31	0.329	22 ± 2	6.7 ± 0.2
HEPES	7.66	0.180	29 ± 3	16.1 ± 0.3
Tris	8.06	0.080	4.0 ± 0.4	5.0 ± 0.1
Bicine	8.46	0.034	7.1 ± 0.7	21.0 ± 0.4
CHES	9.41	0.004	3.8 ± 0.4	96 ± 3

and C.2). Both enzymes were expressed in *E. coli* BL21(DE3), which were lysed, and diluted to a total of 1.8×10^5 fold to mimic droplet conditions. The results for NaPi and Tris buffer are shown in Figure 4.9. The ACP lysate did not show Kemp eliminase activity compared to the lysis buffer and was thus confirmed as a negative control under these conditions. The lysis agent (BugBuster, Merck-Millipore) did not affect the rate of reaction, as previously determined (Appendix C.4). After 90 min of reaction, (53.5 ± 0.4) μM of 1 mM **2a** were turned over by the negative control in NaPi, whereas only (4.7 ± 0.8) μM product appeared in the negative control in Tris buffer. Therefore, the difference in the two buffers remains above 10 fold under realistic reaction conditions. The positive control HG3.17 retained high activity even at 20,000 fold dilution and the reaction saturated after 60 min in Tris buffer.

In summary, reaction conditions were found that allowed clear distinction of a positive and negative control at dilutions equivalent to droplet conditions. It was now time to assess the ability of droplets to retain the reaction product under different conditions.

4.4.2 The Kemp reaction product exchanges between droplets

The system implemented to assess leakage is shown in Figure 4.10. The reaction product **2b** was obtained by converting the substrate **2a** using NaOH in aqueous solution, which was neutralised using the equivalent amount of HCl, and diluted into buffer for each test. Droplets of 70 μm diameter were generated using a chip with two flow-focusing junctions, which allowed the generation of a 1:1 mixture of droplets either containing product or not. A food-dye, tartrazine, was used in both solutions at 3 mM to set the absorbance of buffer below that of the fluoruous oil (see Appendix C.5). The droplets were collected and re-injected for analysis in batches of several thousand after different incubation times. Absorbance was

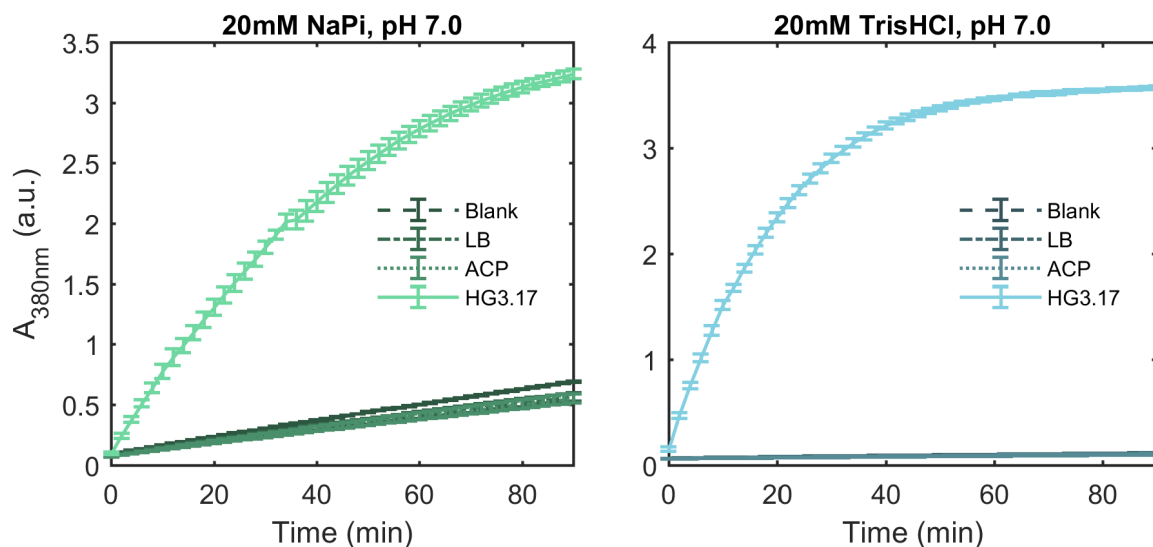


Fig. 4.9 Reaction progress for positive and negative controls of the Kemp Elimination in well-plate format (average of three runs, error bars are the standard deviation). Buffer conditions as indicated. Blank: buffer only control, LB: lysis buffer only, ACP: cell lysate of acidic phosphatase 1 (negative control), HG3.17: cell lysate of positive control HG3.17.

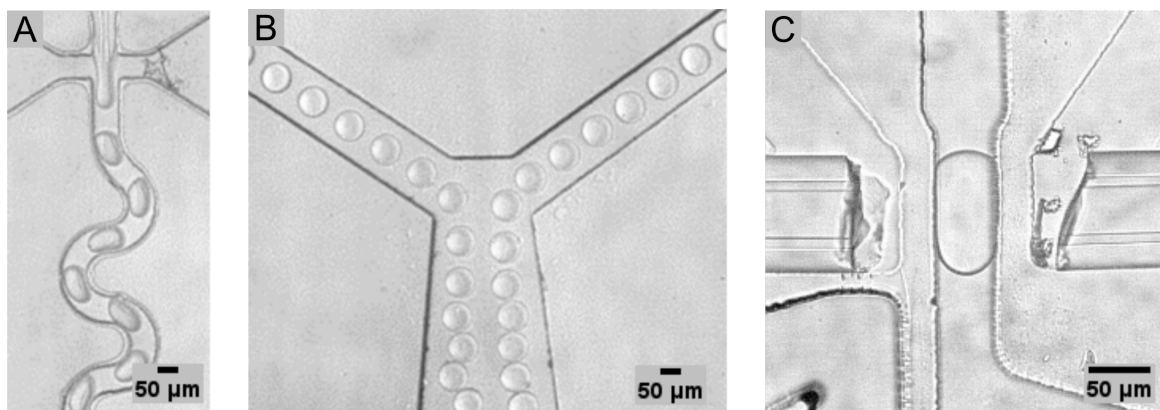


Fig. 4.10 Shown are the microfluidic steps involved in the assessment of product leakage for absorbance-activated droplet sorting. A&B: Droplets are generated in one chip with two flow-focusing junctions to create a 1:1 droplet mixture of droplets made from two different solutions without the need for pipetting. C: These droplets are then re-injected batch-wise into an AADS chip to measure the signal difference over time.

measured at 385 nm in an AADS chip (manufactured according to [61]) under realistic sorting conditions, *i.e.* at 100 Hz.

The signal was measured and analysed using a custom Matlab script as shown in Figure 4.11. The minima of the droplets were determined (which is what the sorting algorithm used as sorting criterion) and a histogram of the minima plotted. A kernel-smoothing dis-

tribution was fitted to the data and the maxima of the two main populations determined. The difference of the maxima gave the signal $\Delta S(t)$ at time t . Usually, the first data ΔS_0 could be acquired 10 min after the end of droplet generation and was used to normalise the signal according to $\Delta S(t)/\Delta S_0$.

Unfortunately, it became evident immediately that compound **2b** exchanged between droplets, *i.e.* it leaked (Figure 4.13). In Tris buffer, half of the initial signal difference was lost 60 min after droplet generation. The decay of the signal was non-linear. A single-exponential fit gave a half-time of 55 min. However, the residual plot revealed systematic deviation indicating complex leakage behaviour similar to what was found in the system of Skihiri *et al.* which consisted of a droplet incubation chamber from which droplets were re-injected into a chip to measure leakage [207].

Seven parameters were varied in attempts to reduce leakage Figure 4.12 and Table 4.2. The conditions were compared to each other based on the time at which half the initial signal was lost, which was determined by a local linear fit of the three nearest data points. Initial experiments were performed at pH 6 in NaPi so as to have a low background reaction. At a surfactant concentration of 2% the two droplet populations immediately equalised even if the droplet generator was directly coupled to the AADS. Reducing the surfactant concentration to 1% under the same conditions led to a signal half time of 25 min. The cavitand β CD was hypothesised to form a complex with **2b** and thus reduce leakage [70]. Indeed, addition of five times excess β CD tripled the half time. Next, it was hypothesised that the uncharged conjugate acid of **2b** could leak more readily than the charged equivalent. At pH 6, 1.2% of the compound would be protonated at any given time, 0.4% at pH 6.5, and 0.1% at pH 7. Indeed, at pH 7 the half time tripled again compared to pH 6.5. This is less than would be explained by the protonation state of **2b**, therefore either both states leak or the effect is due to another mechanism. The surfactant PicoSurf1 (Dolomite) reduced the half time seven times compared to fluorosurfactant-008 (FS-008, Ran Biotechnologies) under otherwise identical conditions. Neat fluorosurfactant oils FC40 and FC70, which are more viscous than HFE7500 thus having a smaller diffusion constant, led to unstable emulsions. A stable 1:1 mixture of HFE7500 and FC70 had a similar half time to neat HFE7500. Finally, the half time was not affected by exchanging NaPi or Tris buffer.

In summary, the influence of the tested parameters was similar to what others found in the study fluorophore leakage [71, 72, 207]. Unfortunately, what was beneficial to product retention, was also promoting the background reaction of the substrate. As Equation 4.3 shows, elevated pH promotes substrate turnover due to catalysis by hydroxide. β CD reduced leakage of product, but was found to catalyse the Kemp reaction and its catalytic activity increased with pH (Appendix Fig. C.6). Another additive, BSA, showed detectable activity in

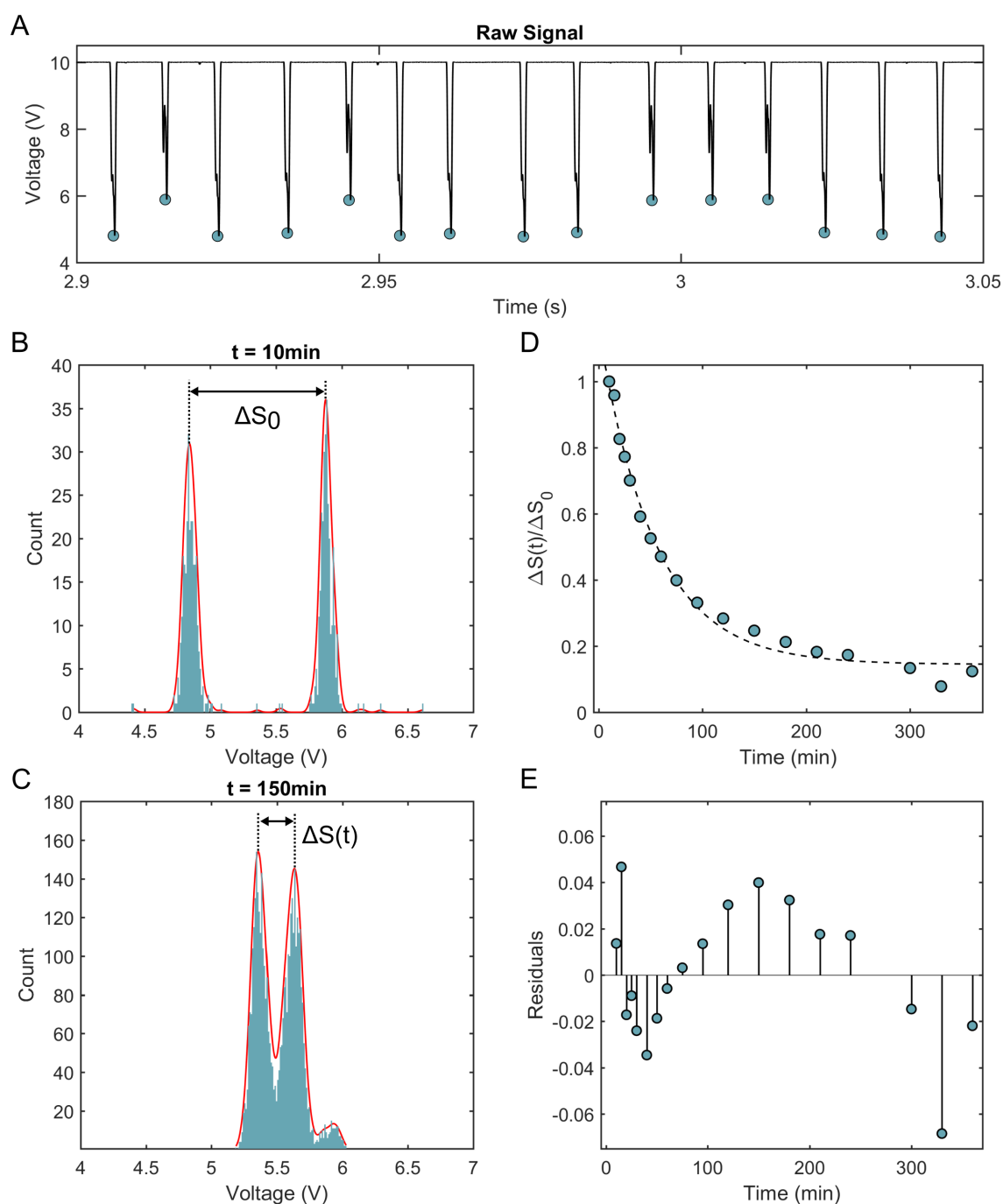


Fig. 4.11 The leakage of Kemp reaction product in droplets. **A**: Droplets that passed through the light path reduced signal due to absorbance. The minimum for each droplet was determined. **B**: The histogram of minima revealed the two intended populations with and without product. A kernel-smoothing distribution was fitted to the data and used to determine the difference in the maxima of the two populations. **C**: Over time, the two main populations moved closer together, indicating the leakage of the Kemp reaction product from filled to empty droplets. **D**: In this experiment, which was performed in 20 mM Tris buffer with 50 mM NaCl, a non-linear decay of the initial signal difference was observed. However, a fitted single-exponential decay (dotted line) did not fit the data well as shown by systematic deviation in the residuals (**E**).

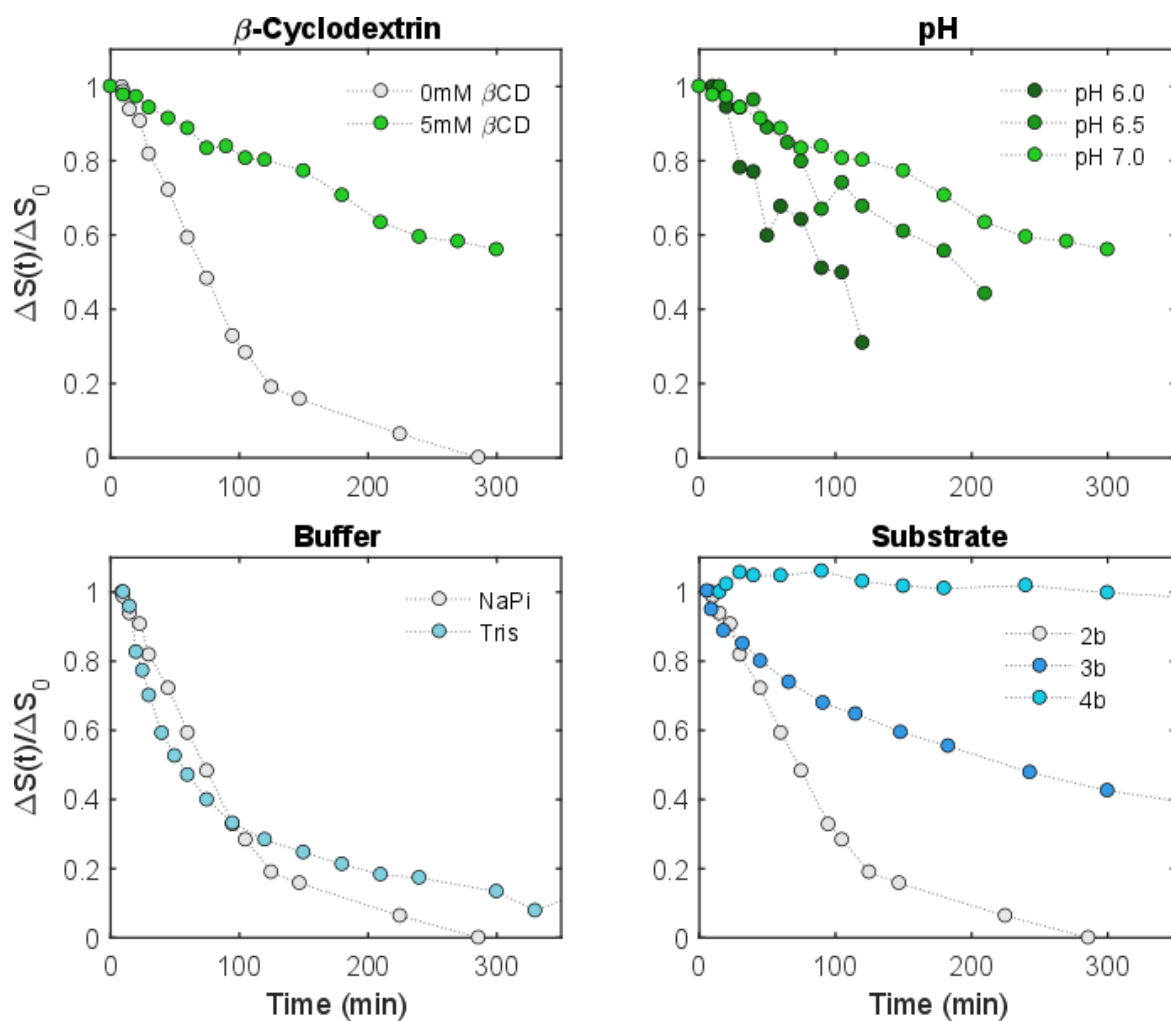


Fig. 4.12 Timecourses of product leakage in a 1:1 mixture of droplets with and without product under different assay conditions. Addition of β CD and increasing the pH reduced leakage. There was no difference between NaPi and Tris buffer. In addition to **2b**), two more Kemp reaction products were tested (structures shown in Figure 4.13), which showed reduced and no leakage, respectively. Table 4.2 lists additional leakage conditions tested.

Table 4.2 The conditions under which leakage of 1 mM Kemp reaction product was tested.

Product	[Surfactant] (w/w)	Surfactant	Oil	Buffer [†]	pH	βCD [‡]	$t_{1/2}^*$ (min)
2b	2%	FS-008	HFE7500	NaPi	6.0		<1
2b	1%	FS-008	HFE7500	NaPi	6.0		25
2b	1%	FS-008	HFE7500	NaPi	6.0	✓	90
2b	1%	FS-008	HFE7500	NaPi	6.5	✓	180
2b	1%	FS-008	HFE7500	NaPi	7.0	✓	320
2b	1%	FS-008	HFE7500/FC70 [§]	NaPi	7.0	✓	290
2b	1%	PicoSurf1	HFE7500	NaPi	7.0	✓	40
2b	0.75%	FS-008	HFE7500	Tris	7.0		60
2b	0.75%	FS-008	HFE7500	NaPi	7.0		70
3b	0.75%	FS-008	HFE7500	NaPi	7.0		230
4b	0.75%	FS-008	HFE7500	NaPi	7.0		>10,000

[†] 40 mM buffer compound and 100 mM NaCl.

[‡] β-cyclodextrin was added (✓) at 5 mM.

^{*} time after which half the original signal was lost.

[§] 1:1 mixture of the two oils.

droplets at 50 μM, which is why it was not tested as an additive (Appendix C.6). Therefore, the assay conditions could not be optimised without a major trade-off between leakage and background activity. Therefore, two approaches were followed to circumvent these limitations:

- First, two additional 1,2-benzisoxazole derivates were synthesized in an attempt to abolish leakage by introducing a charged group.
- Second, a microfluidic device was designed to generate and sort droplets in-line. The advantage of this approach is, that the incubation time of all droplets is very similar thus controlling the maximal leakage time. The limitation is that incubation times beyond 1 to 2 h would be technically difficult to implement, because high back-pressure would destabilise the device.

4.4.3 Modified 1,2-benzisoxazole substrates

The idea for substrate modification was to introduce a permanently charged group to the substrate, which would limit the partitioning into the fluoruous phase but should not itself be able to act as a general base. Two suitable modifications previously used to adapt substrates for droplets are trimethyl ammonium and sulfonate. The syntheses were performed by Dr

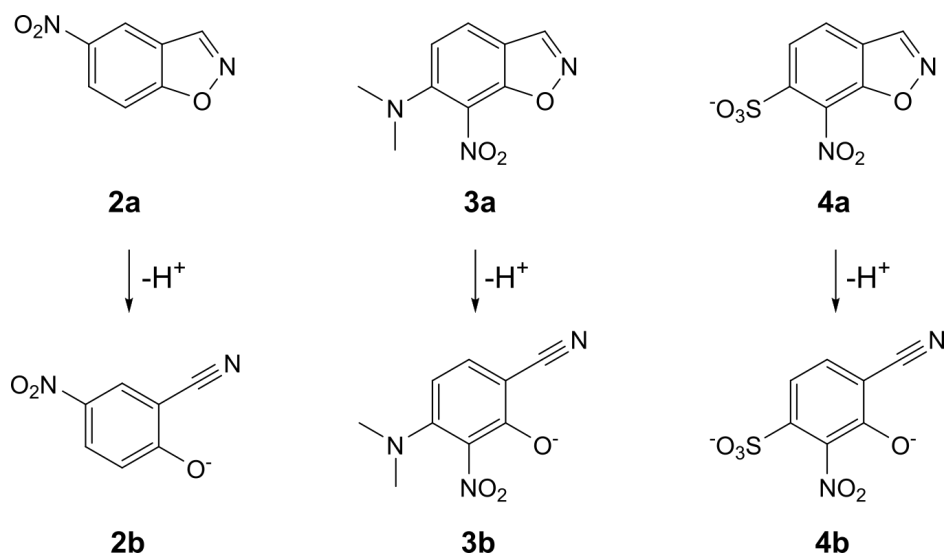


Fig. 4.13 Shown are the Kemp substrates and the respective reaction products, which were tested for leakage in droplets. **3a-4b** courtesy of Dr Josephin Holstein.

Josephin Holstein and yielded 1,2-benzisoxazoles **3a** and **4a**. The original goal for **3a** was to introduce the larger substitution $-\text{N}(\text{CH}_3)\text{C}_2\text{H}_4\text{N}(\text{CH}_3)_3^+$ following the strategy of Obexer *et al.* [47, 79]. However, all attempts to further react the 6-dimethylamino group failed, which left a general base in the substrate itself. Nitration, which was intended to increase the reactivity of substrate and the absorbance of the product, was directed towards the 7 position. Nitration of 6-sulfonyl-1,2-benzisoxazole also occurred in position 7. Substitution of the sulfonate by nitrate was substantial for this substrate and purification by HPLC was necessary to obtain **4a**.

Despite these limitations, **3a** and **4a** were converted under basic conditions to yield the respective Kemp products **3b** and **4b**. The leakage half time of **3b** was extended over 3 fold compared to **2b** under the same conditions. Leakage of **4b** was most strikingly reduced. After seven days of incubation (close to 10^5 min), 65% of the original signal difference remained. While **4b** had excellent product retention, the synthesis of substrate **4a** was not reproducible and numerous attempts to synthesise more of the compound failed. In conclusion, the tested compounds had reduced leakage. However, they were not suitable as alternative substrates.

4.4.4 In-line droplet generation and sorting

Given the rate of product leakage observed, the maximal incubation time was limited to several hours. Because the sorting rate was also in the hours range, this meant that droplets measured at the beginning and end of sorting were not comparable. The idea of the in-line

chip was to fix the incubation time for each individual droplet by linking droplet generation, incubation, and sorting within one microfluidic device².

It connected the droplet generator and the droplet sorter designed by Gielen *et al.* via a delay line with intermittent constrictions [61], see Figure 4.14A. The constrictions were intended to mix the droplets throughout the delay line to avoid broadening of the incubation time [208]. In order to assess product leakage, two flow-focusing junctions were included in the design (Figure 4.14A1). An oil extractor (Figure 4.14A2) removed excess oil needed for droplet generation before the droplets entered 20 loops of incubation chambers. The calculated area of the incubation chambers was 635 mm². With a chamber height of 80 µm and a flow-rate of 1.5 µL min⁻¹, the expected incubation time was 30 min.

It was possible to run the chip stably for several hours. With one droplet generator running, typical flow-rates for stable operation were 30 µL h⁻¹ for each aqueous phase, 120 µL h⁻¹ for the oil phase, and -90 µL h⁻¹ for the oil extractor. This resulted in droplet generation at 50 Hz and, with 85 µm diameter, slightly larger droplets than in the two step (off-line) procedure (the difference in volume was less than 2 fold). The incubation time was found to be longer than expected at 40 and 50 min with the total flow-rate of 1.5 µL min⁻¹ in the incubation chambers. This indicates that under these conditions the flow rate of the droplets in the chambers was mostly determined by the total injection rate of the aqueous phase, while the oil passed the droplets rather than pushing them. However, below a ratio of 1:2 of oil to aqueous flow the dense packing of droplets led to increased merging and at high extraction rates droplets started breaking off at the extractor. After the incubation line, droplets were spaced for the absorbance measurement and sorting. Up to a flow-rate of 20 µL min⁻¹ for the spacing oil, the rate of droplet generation equalled the rate of droplet sorting. Beyond 20 µL min⁻¹, the sorting rate decreased and the droplets started to stall inside the chip and eventually merged.

Two droplet generators were used to test leakage of **2b** at 1 mM using Tris buffer and 0.75% FS-008 surfactant. The flow-rates were 30 µL h⁻¹ for each aqueous phase (one per flow-focusing junction), 90 µL h⁻¹ for each oil phase, and -150 µL h⁻¹ for the oil extractor. The droplet generation rate was 25 Hz in each flow-focusing junction. Under these conditions, only one population of droplets could be observed at the end of the incubation line. This indicated complete equalisation had occurred. It was possible to reduce the surfactant concentration down to 0.1% without compromising the stability of the emulsion³. This enabled discrimination of two droplet populations at the end of the incubation line. Even at

²The concept and initial chips were designed by me. The final design shown here was created with the support of Dr Tomasz Kaminiski

³The minimal concentration of surfactant in the two-step procedure was 0.75% below which droplet merging dominated. This may be due to the higher flow-rates and shear forces acting on the droplets.

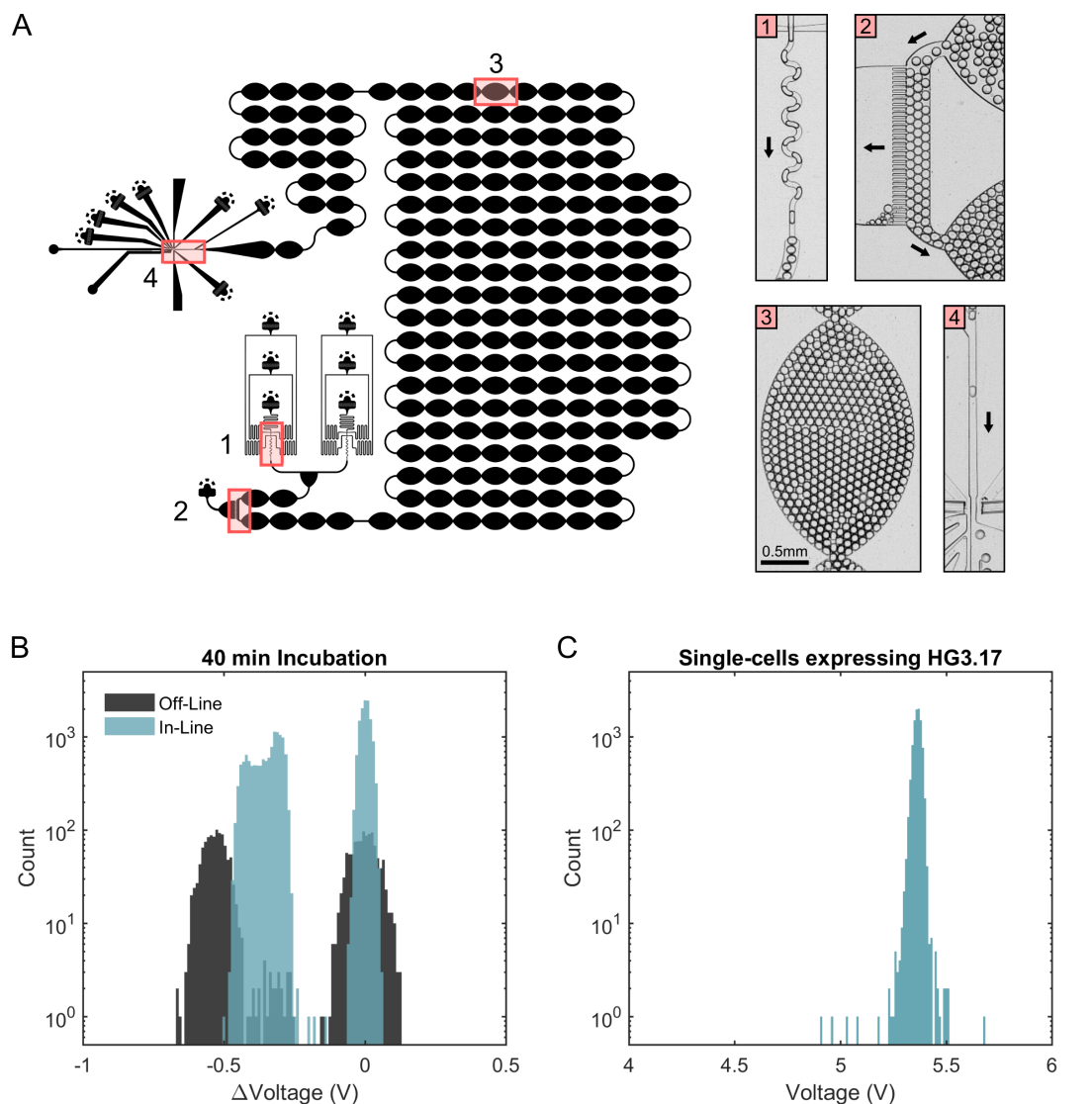


Fig. 4.14 Design and results for the In-Line AADS. **A**: The final design featured two droplet generators (1). One or both could be used to run the chip. The oil extractor (2) removed excess oil to slow the droplets down and pack them for incubation. The incubation chambers (3) featured constrictions to mix droplets to minimise difference in incubation time. Finally, droplets were spaced, measured, and sorted (4). **B**: Leakage of product was found to be increased compared to the two-step (off-line) procedure (off-line 0.75% surfactant, in-line 0.1% surfactant). **C**: Cells having expressed HG3.17 were encapsulated according to a Poisson distribution with λ of 0.35 but no activity could be observed at the end of the incubation line (a second droplet population of several thousand droplets would have been expected).

such low surfactant concentration, the leakage was still about 3-fold stronger in-line (difference of the population means (385 ± 60) mV) than after a comparable incubation time of 40 min in the off-line procedure (difference of population means (600 ± 80) mV), see also Figure 4.14B.

In both set-ups, the complete leakage of product occurred over a time-frame of several hours, whereas completion of the reaction by HG3.17 at single-cell lysate dilutions was reached after about 1 h (Figure 4.9). It was therefore conceivable, that enzymatic activity may be detectable in droplets. Cells having expressed HG3.17 were co-encapsulated with lysis agent at λ of 0.35 and analysed in-line. However, no enzymatic activity was observed as can be seen in Figure 4.14C. There are only five events below background, where thousands would have been expected.

In contrast to this finding, enzymatic activity could be detected in droplets using the off-line system (see the following section and Figure 4.15). There were three possible explanations for this difference: the cells were not encapsulated (*e.g.* due to sedimentation in the syringe), the cells were not lysed efficiently, or the increased leakage in-line prevented a signal from developing. The first was unlikely to be the case, since cell encapsulation at similar flow-rates and at similar time-scales in the esterase assay was efficient (Chapter 3). The lysis agent used was BugBuster (Merck-Millipore) at 0.1x final dilution. An increase of up to 0.5x BugBuster and another lysis agent, polymyxin B at 4 mg mL^{-1} , gave the same results. It was thus concluded that the most likely limitation of the in-line system is the increased rate of product leakage. This may be explained by excess oil flowing past the droplets acting as a leakage sink during incubation. Also, 70% of droplets were empty, potentially acting as product sinks, rather than 50% as in the leakage test.

In summary, a new microfluidic device for on-chip droplet generation, incubation, and sorting based on absorbance was designed and tested. Its operation was stable for several hours. Leakage of the Kemp reaction product **2b** was found to be stronger in-line compared to off-line indicating a difference in leakage dynamics. Enzymatic Kemp activity of HG3.17 could not be detected using this device. However, the chip should be suitable for the screening of enzyme libraries based on assays that do not suffer from leakage of its read-out compounds, *e.g.* those based on formazan dyes [61]. Indeed, a different selection regime becomes accessible, based on measurement during the early (linear) phase of the reaction which is dominated by $k_{\text{cat}}/K_{\text{m}}$. So far, sorting was started 2 h after droplet generation, *i.e.* during the late (plateau) phase of the reaction, which is determined by the amount of enzyme produced and the total turnover number and therefore dominated by changes in expression and stability. For the Kemp elimination, this selection regime was not an option at the mo-

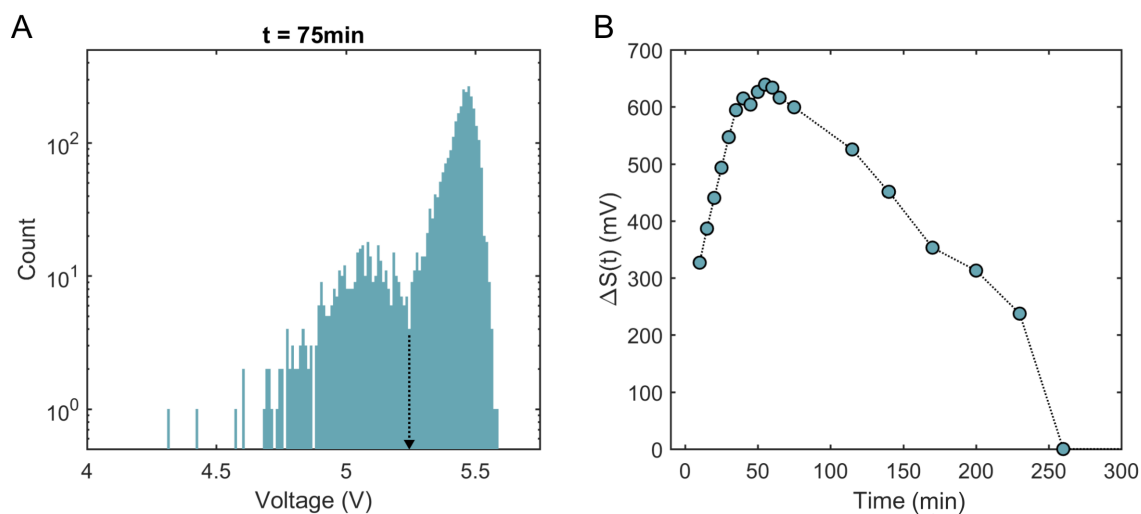


Fig. 4.15 Activity of HG3.17 in droplets. *A*: The number of droplet in the droplet population left of the arrow is 13% of the good agreement with the expected 16% containing HG3.17. *B*: The separation of the two populations increased in the first hour of reaction, indicating enzymatic activity, and decreased thereafter indicating the dominance of leakage.

ment. Therefore, I turned my attention to the two-step procedure again, and tested if the activity of HG3.17 was detectable.

4.4.5 Activity of HG3.17 is detectable in droplets

To test if activity of HG3.17 was detectable in the two-step procedure, cells having expressed HG3.17 and ACP (negative control) were mixed 1:1 and co-encapsulated with substrate and lysis agent at an average droplet occupancy λ of 0.35, *i.e.* about 16% of droplets were expected to contain at least one cell expressing the positive control. Indeed, a droplet population with increased absorbance was observed, as shown in Figure 4.15. Figure 4.15A is a histogram of 3100 droplets of which 400, *i.e.* 13%, are in the population with lower voltage (cut-off indicated by the arrow). This is in good agreement with the expected number of positive droplets and indicates that the second population is due to HG3.17 activity. Both droplet populations were observed 10 min after stopping the droplet generation. The separation between the average signal increased up to 40 min and decreased again after 65 min. The two populations never separated completely. In the well plate assay (Figure 4.9), it took the enzyme about 60 min to reach saturation under similar conditions. Taking this into account, it can be assumed that in the first phase of reaction the rate of product accumulation exceeds the rate of leakage. After 60 min, the situation is reversed and leakage dominates over accumulation. Interestingly, the time after which half the maximal signal difference is lost again, is about 120 min, which is twice that observed in the leakage test under the same conditions. That is

Table 4.3 Results for the enrichment of HG3.17 over N20 using the tributyrin culture plate assay for the phenotypic readout. Halos indicate the presence of N20, no halos indicate the presence of HG3.17.

Ratio	Before Sorting			After Sorting			Enrichment
	Halo	No Halo	ϵ_0^\dagger	Halo	No Halo	ϵ_1	ϵ_1/ϵ_0
1:100	358	9	0.025	11	307	0.965	39
1:1000	47	0	0.001	3	194	0.985	985

[†] ϵ is the ratio of non-halo forming colonies over halo forming colonies. ϵ_0 for the 1:1000 dilution is assumed.

the leakage was apparently reduced in the cell lysate assay. However, faster leakage would have been expected because almost 9 in 10 droplets acted as potential leakage sinks. This indicates that conversion of substrate continues beyond 60 min possibly due to the transfer of substrate from negative towards positive droplets. In summary, these observations strongly indicated that activity of HG3.17 was transiently detectable in 70 μm droplets from single-cell lysates. To prove that this was indeed the case, an enrichment of HG3.17 over a negative control was performed.

4.4.6 Enrichment of HG3.17 using AADS is efficient

To assess the enrichment of HG3.17, a simple culture plate assay was established by using the esterase N20 (see Chapter 3) instead of ACP as the negative control. Colonies harbouring the plasmid pHAT_N20 developed a clear zone (halo) around them if growing on 1% tributyrin plates. The halo appeared without the need for cell lysis, in the absence of expression induction, and in both the BL21(DE3) and the commercial *E. coli* 10G strains of *E. coli*. Therefore, either strain could be plated onto tributyrin plates to measure the ratio of cells harbouring the genes for N20 and HG3.17 respectively. This was convenient, because the former strain was needed to overexpress HG3.17 for the droplet assay, whereas the latter was needed to recover plasmids after droplet sorting. For the purpose of this enrichment, the HG3.17 gene was cloned into the pHAT vector.

Two enrichments were performed (see Figure 4.16 and Table 4.3). In the first enrichment, cells having expressed HG3.17 and N20 were mixed 1:100 and encapsulated at λ of 0.35. Over 60 min, 1.5×10^5 droplets were sorted of which 150 were collected. Half of the recovered DNA was transformed and 318 colonies obtained, *i.e.* about 4 colonies per collected droplet. The percentage of colonies changed from 2.5% without halos prior to sorting to 96.5% without halos after sorting indicating 39 \times enrichment. The maximum possible en-

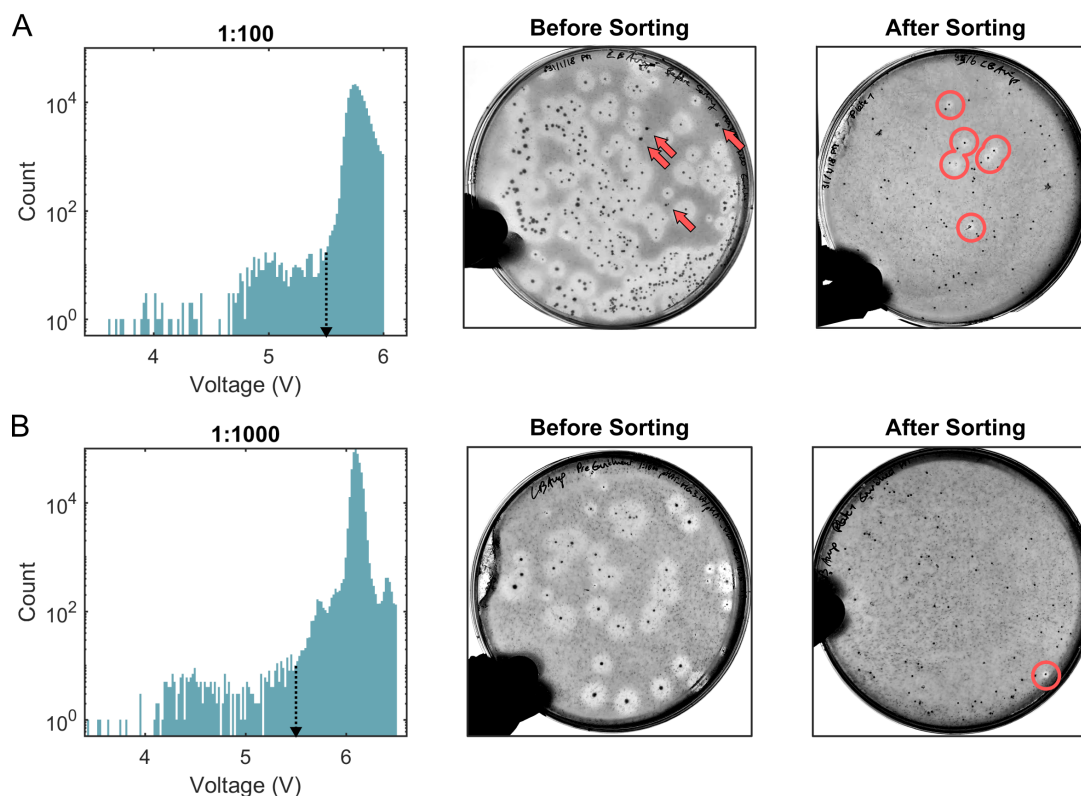


Fig. 4.16 Enrichment of Kemp eliminase HG3.17 over esterase N20 at ratios of 1:100 (A) and 1:1000 (B). Shown are the respective droplet histograms and tributyrin-containing culture plates from before and after the enrichment. The disappearance of the halos indicates enrichment of HG3.17. The dotted arrow indicates the sorting threshold, red arrows indicate colonies without a halo before sorting, red circles indicate colonies with halos after sorting.

richment from 2.5% would have been 40×, *i.e.* the sorting was 97% efficient. Ten colonies were picked for colony PCR with gene-specific primers to confirm the phenotypic readout. The result was as expected: all of the colonies without halo were positive for the HG3.17 gene but not the N20 gene and *vice versa* (Figure 4.17).

A second enrichment with a starting ratio of 1:1000 was performed and similar results obtained. This time, 4.5×10^5 droplets were sorted over 120 min and 130 droplets were collected. Again, about 4 colonies per sorted droplet were obtained. The starting ratio could not be confirmed on culture plate, because too few colonies grew (all of which formed halos). Assuming the starting ratio was as intended, the enrichment was close to 99% efficient.

In summary, it was shown that HG3.17 activity was detectable in single-cell lysates using 70 µm droplets. Efficient enrichment based on this activity was possible in a time-frame of at least 2 h. In principle, enrichment of a well-expressed enzyme with high Kemp eliminase activity was now possible.

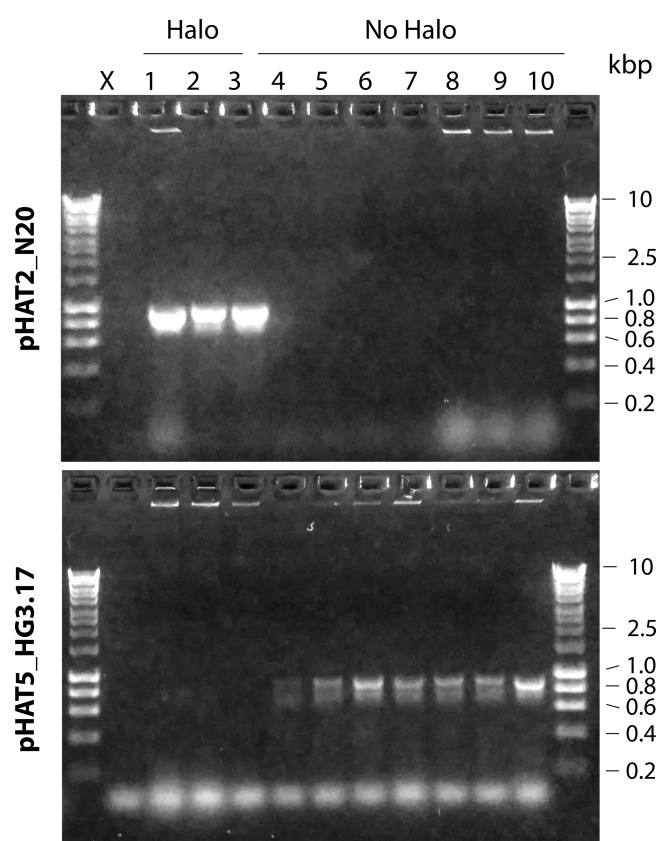


Fig. 4.17 The phenotypic readout of the culture plate assay was confirmed via colony PCR. Ten colonies from after the 1:100 enrichment were picked. Two PCR reactions performed on each using primers specific to the pHAT2_N20 and pHAT5_HG3.17 plasmids respectively. In each case, the phenotype predicted the genotype correctly (X = negative control).

4.4.7 No activity observed in functional metagenomic screening

Cells were transformed and grown over two days on plates to allow for gene expression and then encapsulated using the same conditions as in the enrichments. This time, 3×10^5 droplets were analysed. There were events with absorbance above background. However, all of these were artifacts such as merged droplets or dust (Figure 4.18). No colonies could be recovered from the collected droplets. Therefore, the assay using substrate **2a** appeared not to be sensitive enough for functional metagenomic screening. In the next section, the suitability of this assay to screen mutagenic libraries of HG3.17 at unprecedented throughput was explored instead. The ability to perform a functional metagenomic screen for Kemp eliminases was further investigated using a novel fluorogenic substrate, first reported here, in Chapter 5.

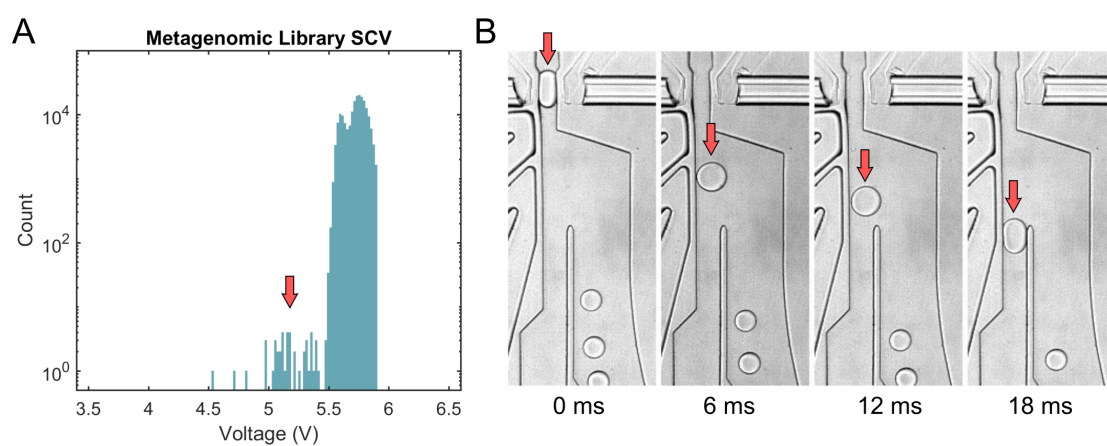


Fig. 4.18 No genuine activity was observed when screening cells transformed with the metagenomic SCV library. All events with a voltage below background were artifacts (*e.g.* large droplets) as indicated in A and B by the red arrow. No colonies could be recovered from the collected droplets.

4.5 Directed evolution of HG3.17 using AADS

The absorbance-based Kemp elimination assay was shown to be efficient at enriching HG3.17 wild-type enzyme over a negative control. Therefore, I explored the utility of this assay in screening large mutagenic libraries of Kemp eliminases. The screening of Kemp eliminases has been limited by the throughput of 96-well plate assays. The enzyme HG3.17 itself was evolved from HG3 using substrate **2a**. Each round, ten 96-well plates were screened, a practical limitation putting the throughput at $>10^3$ [16]. The total number of variants screened by Blomberg *et al.* was in the range of 10^4 – a number which using AADS at a $\lambda = 0.35$ could be screened in less than 10 min. Therefore, the utility of pre-screening libraries of HG3.17 using the droplet method was tested.

Three different libraries were prepared. The first introduced random amino acid substitutions by point mutation (changing one base at a time in the sequence); the second consisted of deletions of a triplet base pair at a random position in the gene (deletion of one amino acid per sequence); and the third contained insertions of an NNN triplet basepair (N being any base, adding one amino acid per gene). Figure 4.19 shows how changes in the nucleotide sequence in these libraries can affect the protein sequence.

Together point substitutions, insertions and deletions (InDels) account for the majority of changes in the evolution of proteins [209]. Substitutions occur more frequently with the

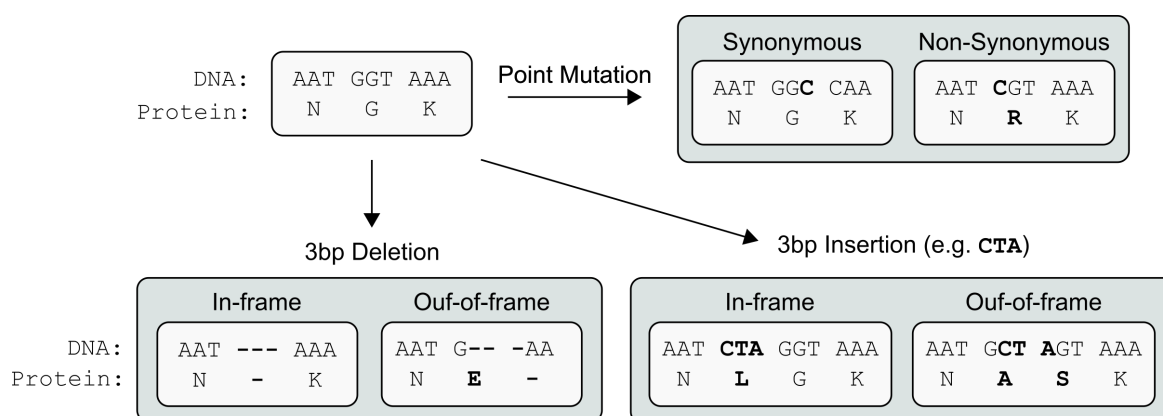


Fig. 4.19 Three libraries of HG3.17 were prepared. The first introduced point mutations changing one base-pair at a time. This method can lead to synonymous mutations (conserving the protein sequence due to the redundancy of the genetic code) and non-synonymous mutations (change of the protein sequence). The method used to generate 3bp InDels allowed the introduction or removal of 3 base-pairs anywhere along the sequence. Thus, in the absence of bias, 1/3 are in-frame and 2/3 out-of-frame. Out-of-frame InDels are likely to cause an amino-acid mutation adjacent to the InDel as shown by the examples. The construction of the libraries is explained in Figure 4.20.

ratio to InDels being 1:20 in the genomes of bacteria [210]. This drops to 1:5 in non-coding regions and increases to 1:40 in coding regions, indicating that InDels occur frequently, but are under stricter selection compared to substitutions [210]. Point substitutions change the side chain of one amino acid, often having locally confined effects. In contrast, InDels change the length of the protein backbone, requiring repositioning of both the backbone and side chains. Due to this more drastic impact on the structure, InDels may be more deleterious to protein function, rationalising the increased selection pressure. Accordingly, substitutions have been studied much more extensively than InDels both in general [210–212]. Yet, InDels occur frequently, [210, 211], and have been shown to be associated with crucial steps in functional divergence of proteins [180, 211, 213, 214].

InDel libraries of enzymes have a lower proportion of active variants not just because of the larger impact on the protein's function, but also because only InDels of $3n$ base pairs avoid highly deleterious frame-shifts. Randomised approaches not controlling for the number of base-pairs inserted or deleted cause >66% frame-shifts (for example [215]). Here, I used an unpublished method developed by Dr Stephane Emond dubbed Transposition-based Random Insertion And Deletion mutagenesis (TRIAD, unpublished). This method produces a majority of in-frame InDels based on engineered transposons. The cloning strategy is outlined in Figure 4.20 and the detailed procedures described in the Methods.

In the following, I first created a substitution library to evaluate the ability to enrich active HG3.17 library variants using AADS and then screened deletion and insertion libraries, respectively.

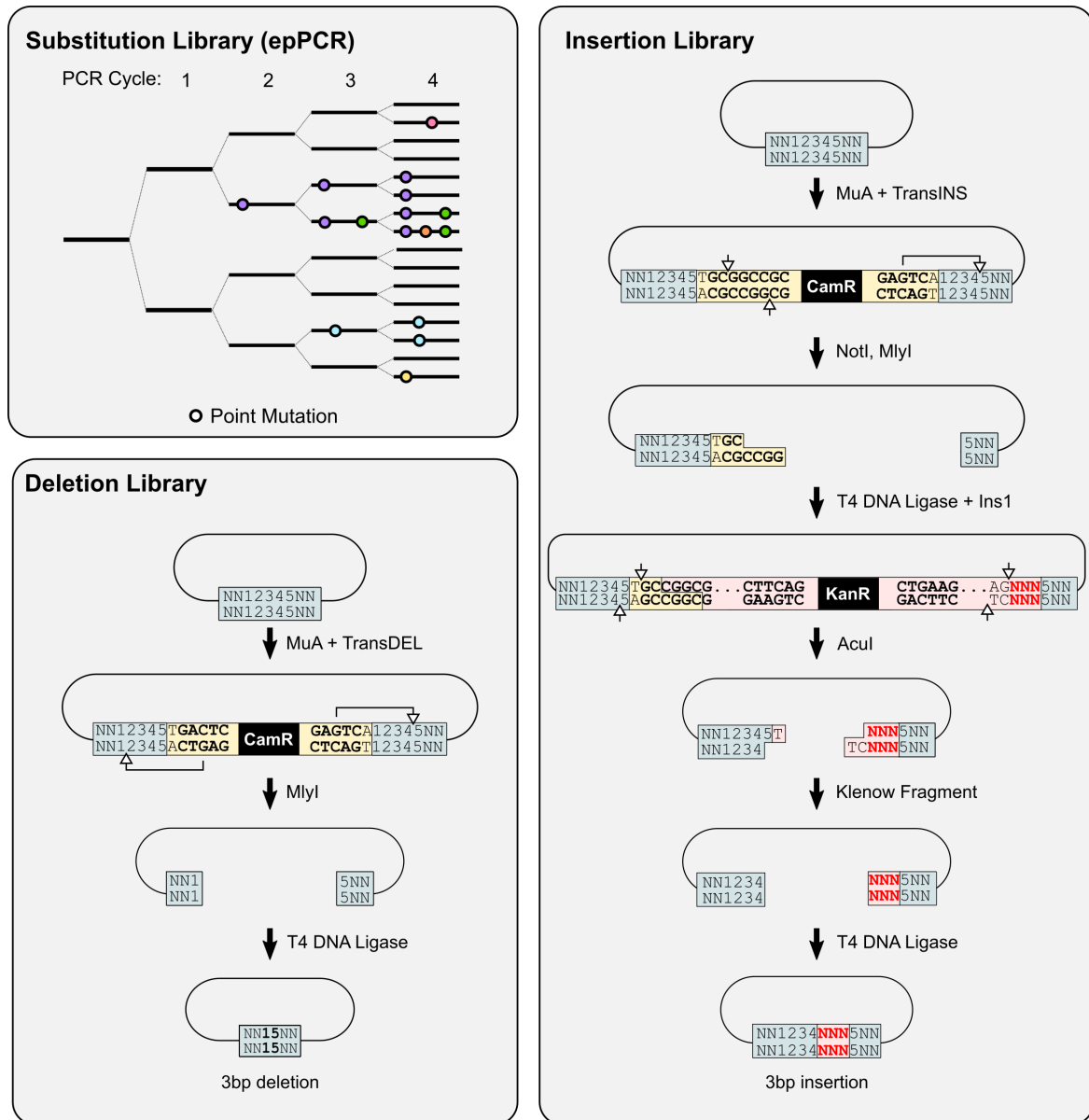


Fig. 4.20 Three libraries of HG3.17 were generated. The first library was created using error-prone PCR (epPCR), which introduced point mutations into the gene based on an error-prone DNA polymerase. The other two libraries were generated using transposable elements to delete or introduce three base pairs at random positions in the target gene (unpublished method developed by Dr Stephane Emond).

4.5.1 The screening workflow

Every library was screened following the same workflow shown in Figure 4.21. First, the library was transformed into *E. coli* BL21-Gold(DE3) (Agilent), which were plated onto LB agar plates. After overnight incubation at 37 °C to allow the colonies to grow, the plates were scraped and the resulting cell suspension used to start two liquid cultures. These were induced using isopropyl β -D-1-thiogalactopyranoside (IPTG) for protein expression at 20 °C. After overnight expression, these cells were used to generate droplets using the same reaction conditions as in the enrichment experiment of wild-type HG3.17 against esterase N20 (Section 4.4.6). Immediately after droplet generation, AADS was used to sort the droplets at 100 Hz for up to three hours. After sorting, the DNA was recovered from droplets using a highly competent commercial strain of *E. coli* (E cloni 10G Elite, Lucigen, $>10^{10}$ cfu/ μ g). The DNA was then extracted and transformed into the expression strain (BL21-Gold(DE3), Agilent, 10^7 cfu/ μ g) and individual colonies picked to inoculate 96-well plates. Cell lysate assays were performed to compare the distribution of enzymatic activity in the library before and after droplet sorting and to pick individual variants for further characterisation.

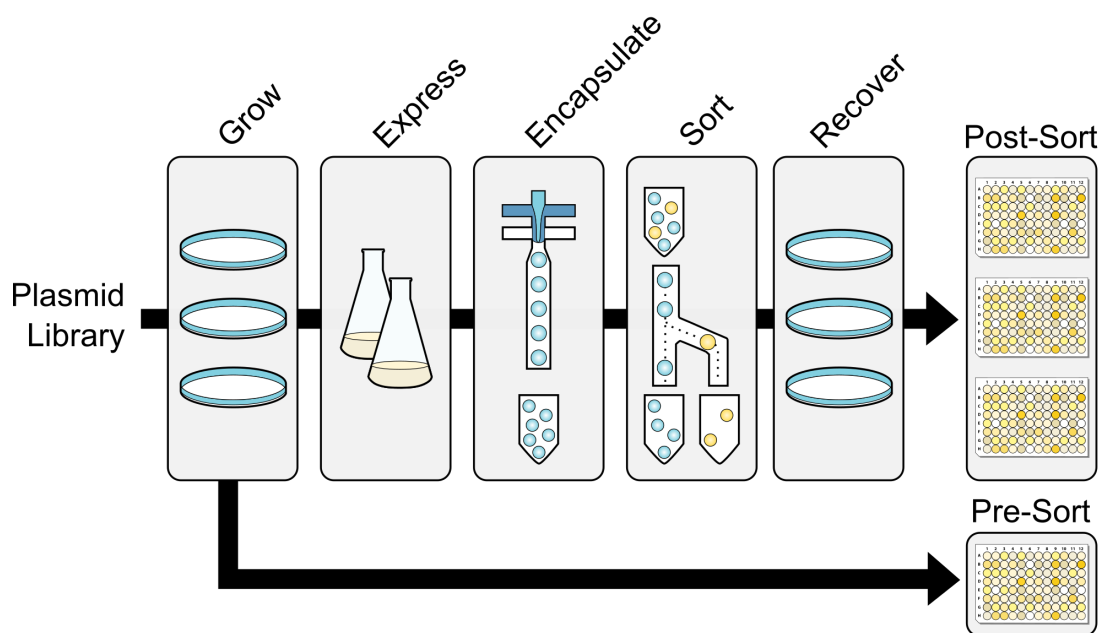


Fig. 4.21 Workflow used to screen the HG3.17 libraries using substrate **2a** and AADS. Droplet screening was performed using the same reaction conditions as in the HG3.17 enrichment experiment (Section 4.4.6). Cell lysate assays were performed to compare the distribution of activity in the libraries before and after droplet screening.

4.5.2 Directed evolution of HG3.17 using substitutions or InDels

Screening of an epPCR library of HG3.17

Error-prone PCR (epPCR) introduces point mutations into a gene by means of a low-fidelity polymerase and variation of the reaction conditions [216]. This is a widely used approach in protein engineering and directed evolution [9, 217] and was previously used in the directed evolution of HG3.17 [16]. Error-rates can range between 1 and 20×10^{-3} depending on the reaction conditions [217]. While epPCR provides access to large substitution libraries, it is important to note the full theoretical sequence diversity is not accessible by this method. Due to the low frequency of the mutations, usually only one base is modified in any one codon. Therefore, not every amino acid can be mutated to any other. For example, a change from valine (GUN) to threonine (ACN) requires both of the first two base-pairs to mutate. The difference between the maximum diversity and the accessible diversity is sequence-dependent and was calculated for HG3.17 using the program PEDEL-AA [218]. With a length of 307 amino acids (His-tag excluded), the maximum library size for one amino acid mutation per protein is $307 \times 19 = 5833$. If allowing only 1 base-pair substitution per codon the accessible number of single amino acid substitutions is 1843. For two amino acid mutations per protein, the maximum number of protein variants is 1.7×10^7 and the accessible variants by single base pair mutations 1.7×10^6 . The actual number of variants contained in a library for any number of amino acid mutations is between the two extremes and depends on the average mutation frequency, bias and library size.

Generation of the epPCR library Here, the GeneMorph II kit (Agilent) was used (detailed reaction conditions in Section 7.2.4). The mutational frequency, ligation and transformation efficiency can vary between different library preparations. Therefore, four libraries were constructed based on independent epPCR reactions. Two epPCR reactions were performed each at 25 and 30 PCR cycles and the results are shown in Table 4.4.

Table 4.4 Substitution libraries generated by epPCR.

Sample	Cycles	Transformants	Insert	Library Size	Mutations/gene
25.1	25	30×10^3	1/10	3×10^3	-
25.2	25	300×10^3	4/10	120×10^3	3.5
30.1	30	40×10^3	5/10	20×10^3	1.8
30.2	30	30×10^3	4/10	12×10^3	1.8
epPCR			9/9	152×10^3	2.7

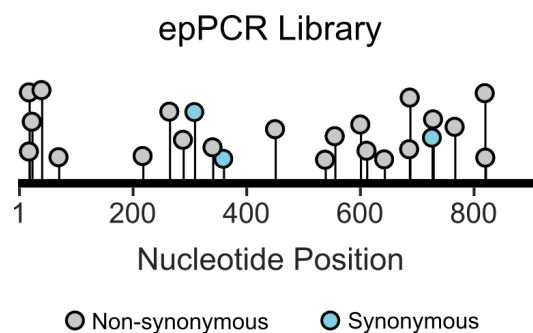


Fig. 4.22 The circles indicate the positions of nucleotide mutations found in 9 randomly picked variants from the naive epPCR library (on average 2.7 mutations/variant). At the amino acid level there were 2.3 substitutions/variant on average.

Ten variants were sequenced from each library. The library size was estimated by multiplying the total number of variants obtained by the proportion of variants containing vector with insert. Sample 25.1 resulted in a particularly small library and was discarded. The largest library was obtained from sample 25.2 with an average of 3.5 mutations per gene. Because all libraries had a high proportion of empty vector, a portion of libraries 25.2, 30.1 and 30.2 was digested using restriction enzyme *EcoRI* (only present in vector without insert), the purified product mixed at ratios representing the original library size and freshly transformed. All of the sequenced variants of this final library (epPCR) contained insert and the estimated diversity was on the order of 10^5 . The mutational bias in the library was estimated using the program Mutanalyzer [219]. The ratio of transitions to transversions was 0.6 (close to the ideal of 0.5) and the ratio of $AT \rightarrow GC$ to $GC \rightarrow AT$ was 0.4 (similar to 0.6 reported by the manufacturer). Mutanalyzer fits a Poisson distribution to the input mutations and estimated a $\lambda_{mut} = 2.8$, meaning the library is expected to contain about 6% wild-type sequence and about 20% each with one, two and three base pair mutations, respectively. The mutations observed in the randomly picked variants were distributed across the whole gene (Figure 4.22).

In the following, I report on two screening campaigns of this library. In the first campaign it was tested how sorting stringency (top fraction of droplets collected) influences the enrichment of activity. During the second campaign two subsequent rounds of enrichment were performed to test whether improved variants can be further enriched.

First Sorting Campaign The epPCR library was transformed into chemically competent *E. coli* BL21-Gold(DE3) cells resulting in a library of 2.5×10^4 clones⁴. While this was only a

⁴Construction of the final library was performed using highly electro-competent *E. coli* cells (E cloni 10G Elite, Lucigen) covering the estimated size of 1.5×10^5 at least 10 fold. Electro-competent BL21-Gold(DE3) cells were subsequently prepared (Methods 7.2.2) to achieve higher transformation efficiencies for library expression and droplet sorting.

subset of the epPCR library, it was 10 times larger than any library screened before in a single round for the Kemp elimination. This library was used to investigate the influence of sorting stringency on activity enrichment.

Droplet screening reduces library size by enriching the top-performing variants in the library. The cut-off point is determined by the sorting threshold of the sort and determines the ratio of the input library size N_{lib} to the output library size N_{out} . This ratio is not only influenced by the sorting threshold, but also by the coverage C of the library:

$$C = \frac{N_{\text{cells}}}{N_{\text{lib}}} = \frac{N_{\text{tot}} * \lambda}{N_{\text{lib}}} \quad (4.5)$$

With N_{tot} being the total number of droplets screened and λ the average droplet occupancy of the sample. If a library is undersampled ($C \ll 1$), then the output library size will approximately equal the number of collected droplets ($N_{\text{coll}} \approx \text{collected cells}$), because the likelihood of collecting the same library member twice is low. If a library is oversampled ($C \gg 1$), then the output library size will be approximately the number of collected droplets divided by the coverage, assuming the same library member will give a similar signal every time it is measured:

$$C \ll 1 : N_{\text{out}} \approx N_{\text{coll}} \quad (4.6)$$

$$C \gg 1 : N_{\text{out}} \approx \frac{N_{\text{coll}}}{C} = \frac{N_{\text{coll}} * N_{\text{lib}}}{N_{\text{tot}} * \lambda} \quad (4.7)$$

By reducing the droplet sorting threshold (higher absorbance) the collection frequency will be reduced. More droplets will need to be screened, increasing C , but fewer droplets can be collected due to the time limitation of the assay, both effects increasing stringency. Higher stringency is advantageous because the smaller output library can be fully re-screened in the 96-well plate assay (e.g. an output of 10 can easily be oversampled in a single plate). However, higher stringency also increase the chance of erroneous sorting events occurring due to the prolonged sorting time, introducing a stringency trade-off.

To investigate the effect of sorting stringency in this assay, the epPCR library was sorted using two different thresholds. Both sorts were performed on the same day. The droplet samples were each produced immediately prior to sorting and using the same cell suspension (stored on ice between the two sorts).

The results are shown in Figure 4.23 and Table 4.5. Both sorts showed a similar signal distribution with a main peak just above 6 V and trailing off approximately exponentially towards lower voltages. Almost no droplets had a signal below 4.5 V. The first sort was per-

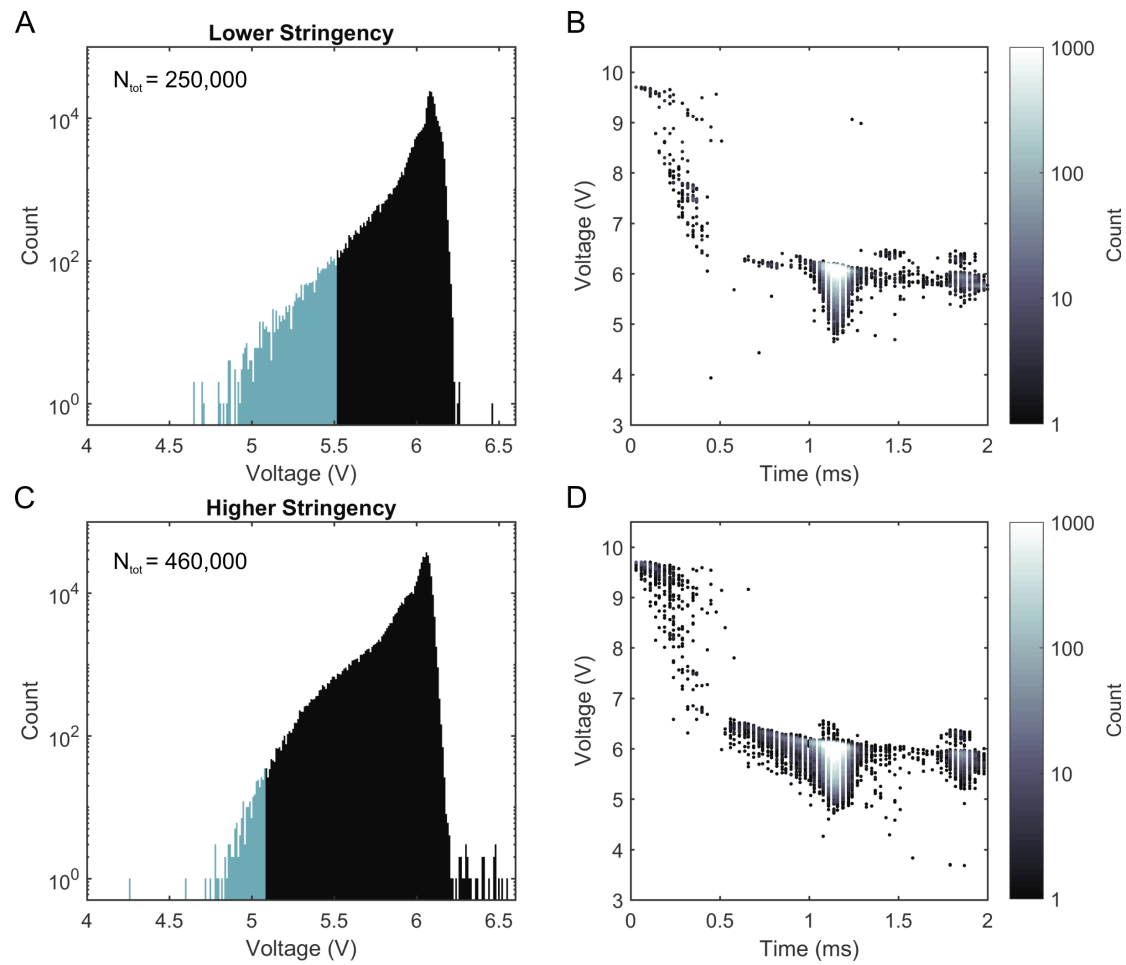


Fig. 4.23 Histograms of two droplet sorts of the epPCR library of HG3.17 using AADS. The first sort was performed at a lower stringency (top 0.8% of droplets) and the second at higher stringency (top 0.06% of droplets). Droplets in the blue range were collected.

formed at a lower stringency with a sorting threshold of 5.5 V resulting in the collection of 1 in 100 droplets. Shifting the threshold to 5.1 V reduced the collection frequency to 5×10^{-4} . The output library sizes were estimated to differ about 15 fold.

Table 4.5 Number of clones analysed for the two droplet sorts of the epPCR library of Kemp eliminase HG3.17.

Stringency	Threshold [†] (mV)	Cells Analysed $N_{\text{tot}} * \lambda$	Coverage C	Droplets Collected N_{coll}	Output Library N_{out}
Lower	580	9×10^4	3.5	2.0×10^3	~ 600
Higher	960	16×10^4	6.4	2.6×10^2	~ 40

[†] difference to the mode of the signal distribution.

Upon DNA recovery, the lower stringency sample yielded 4 colonies per collected droplet and the higher stringency sample yielded 3 colonies per droplet. Individual colonies were picked from before droplet screening and for each sample and grown in deep 96-well plates. After protein expression in 500 μ L medium, the cells were pelleted and lysed in 100 μ L lysis buffer. The lysis solution was diluted step-wise to 2×10^4 times. Assuming an average OD₆₀₀ of 3 to 3.5 for each well, the dilution of the cytosol was about 10 fold higher compared to droplets. The results for the cell lysate assay are shown in Figure 4.24.

The reaction rates during the first 5 min were calculated for each well and normalised to the HG3.17 (positive) and N20 (negative) controls. Almost half of the clones tested for the naive library activity fell below 10% cell lysate activity compared to the HG3.17 control. This fraction was reduced by more than 2 fold to 18% in the lower stringency sample, but to only 33% in the higher stringency sample. Thus, this experiment established that sorting with a lower threshold was more efficient at removing inactive variants from the library. As explained above, lower enrichment of activity at high stringency can be explained by the prolonged sorting time, which increases the number of erroneously collected droplets due to low-frequency events such as small dust particles disturbing the flow in the sorting chip.

No clones with activity more than 2 fold above HG3.17 were observed. There were four clones with activity 1.5 fold above wild type HG3.17, which were selected for further analysis. They were re-grown in 3 mL medium and the protein expressed overnight. The OD₆₀₀ was adjusted to 0.5 to account for differences in the growth rate and 1.8 mL of each sample was spun down to repeat the cell lysate assay, plasmids were extracted from the remainder of the sample for sequencing. The result is shown in Figure 4.26: as in the previous assay, all of the selected variants showed higher activity compared to the control (ranging from 1.2 to 1.6 fold depending on substrate concentration).

Sequencing revealed that clones 1D6, 3F10 and 4F11 were all equal to the HG3.17 wild-type sequence. Clone 4A11, which was the most active in both cell lysate assays, had two synonymous mutations (C90T, T693C). While all four clones were therefore identical to HG3.17 at the amino acid level, the small increase in activity was reproduced in the repeat experiment. This may be explained by synonymous mutations affecting protein expression levels to a small extent. The C90T mutation in 4A11 changed an ACC codon to ACT (threonine). *Boel et al.* found the first codon to be associated with reduced expression and the latter with improved expression in a large-scale protein expression study [220]. The T693C mutation changed a GGT codon to GGC (glycine), which was found to be near neutral in the same study. Factors put forward to explain changes in expression include effects on mRNA folding, stability and translation rates, but remain a matter of debate [220]. The increased activity

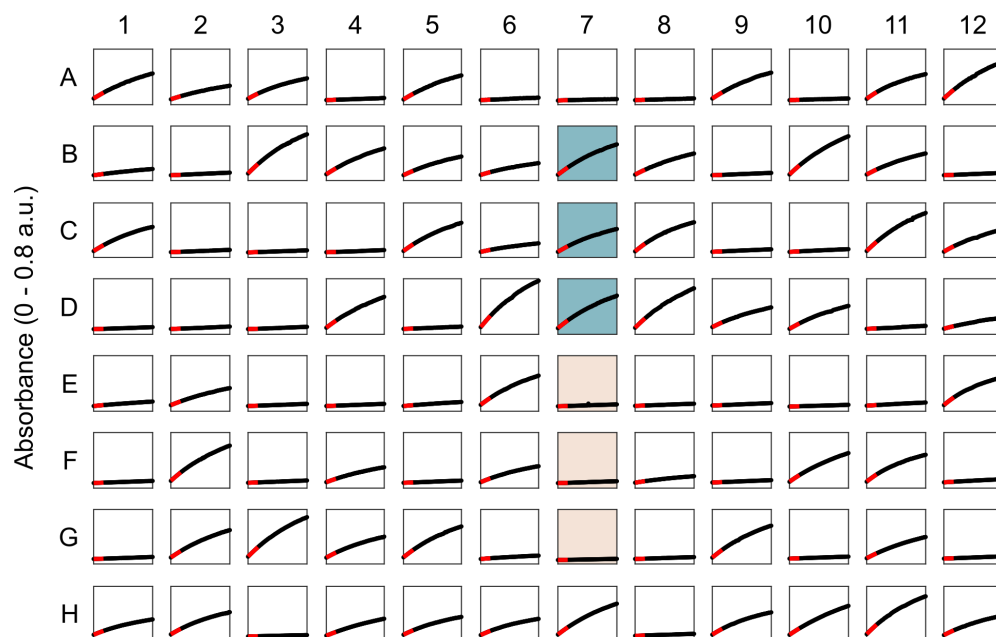
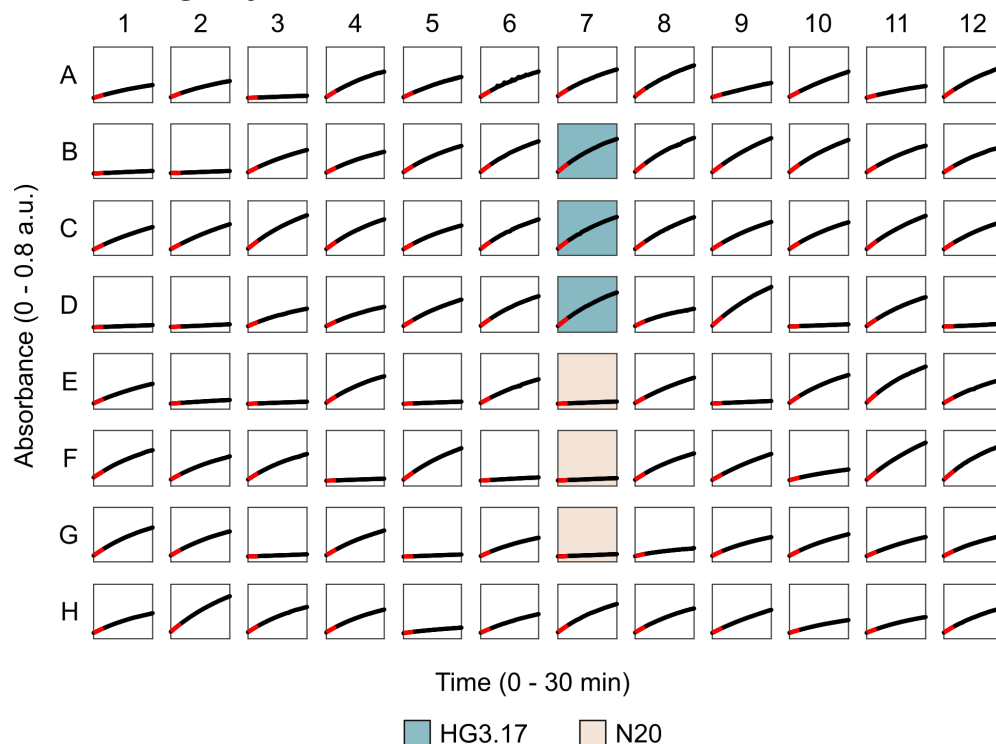
Naive epPCR Library**Lower Stringency Sort**

Fig. 4.24 Cell lysate plate assays of the naive epPCR library (top plate) and after sorting at lower stringency (bottom plate). Reactions were started by adding 1 mM **2a** to diluted cell lysate (2×10^4 fold) and monitored at 380 nm (using UV-transparent 96-well plates). Buffer: 20 mM TrisHCl pH 7, 50 mM NaCl, 10% v/v MeOH.

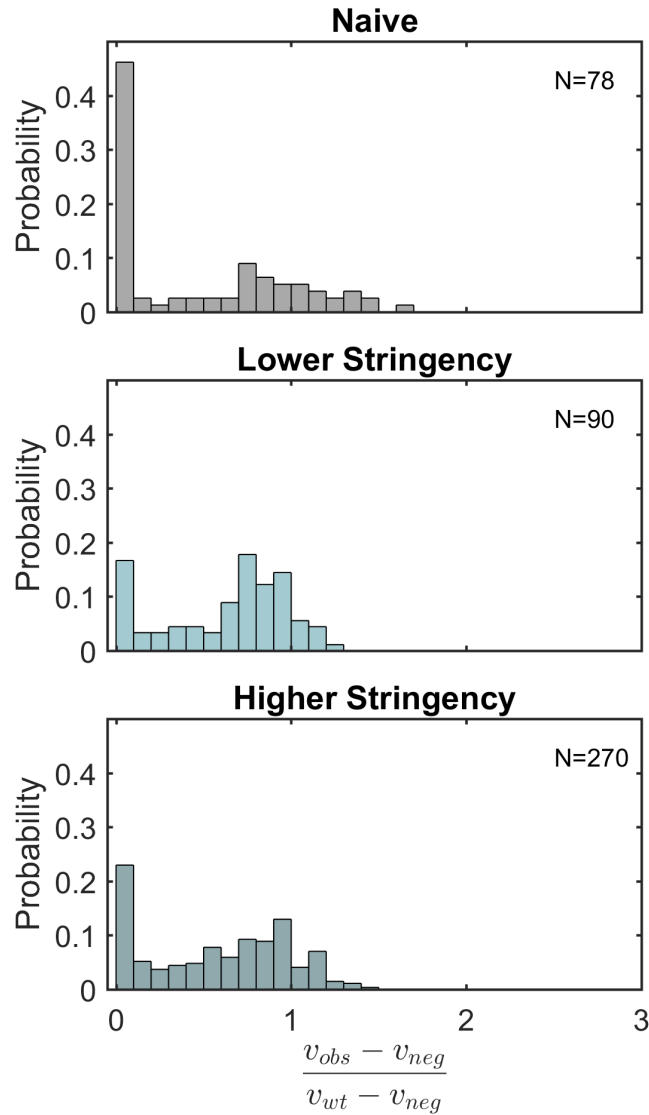


Fig. 4.25 The rate during the initial 5 min of reaction was obtained by linear regression (red lines in Figure 4.24) and normalised with respect to the controls. The resulting histograms show the distribution of cell lysate activity. Shown are the probabilities to account for the differing number of samples N , bins are in 10% steps. Droplet selection at lower stringency was more successful at removing low activity variants (reduction from 46% to 18%) compared to higher stringency (33% remaining).

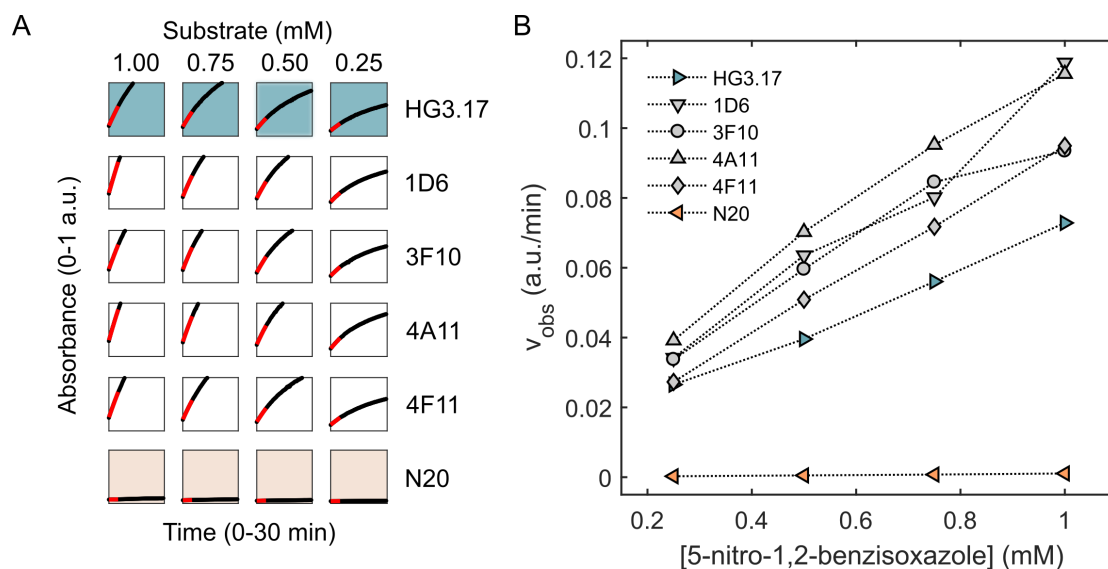


Fig. 4.26 A: Cell lysate assay of four selected variants at different substrate concentrations (OD_{600} was equalised prior to cell lysis). The red lines are a linear fit of the first 5 min of the reaction with slope v_{obs} . B: All four selected variants showed increased activity compared to HG3.17 (substrate concentration during both droplet and plate screening was 1 mM).

in the other clones could be explained by a mutation elsewhere on the plasmid (*e.g.* affecting the plasmid copy number) or in the host genome.

In summary, droplet screening for Kemp eliminase enriched active library members and was more effective in doing so at the lower stringency. While the library screened here did not yield any strongly improved variants, it was shown that the plate assay was sensitive towards small improvements in cell lysate activity. In the next section this approach was further explored by screening the epPCR library over two rounds.

Second Sorting Campaign Given that the higher stringency sort was less efficient, the alternative route of sorting the library over two rounds using two steps at the lower stringency was tested in this second sorting campaign. The epPCR library was transformed again, this time into electro-competent BL21-Gold(DE3) cells, yielding $\sim 1.2 \times 10^5$ transformants.

This library was sorted twice, with the DNA recovered from the first round used as the input for the second round. The results are shown in Table 4.6 and Figure 4.27. In the first round, sorting was performed using even lower stringency compared to the above sorts, with an aim to reduce the library size from $\sim 10^5$ to $\sim 10^4$. The second round was performed using 0.75 mM substrate as an alternative way to increase stringency. During the second sort a defined shoulder to the left of the main peak was observed. The fraction of events below the intervening minimum (at 5.93 V) was 0.35 matching the λ used to encapsulate cells. This

indicated that after the first round of sorting, all library members had detectable activity in droplets above a certain minimum.

Table 4.6 Number of clones analysed during each round of enrichment of the epPCR library of Kemp eliminase HG3.17.

Sample	Threshold [†] (mV)	Cells Analysed $N_{\text{tot}} * \lambda$	Coverage C	Droplets Collected N_{coll}	Output Library N_{out}
Round 1	340	3.5×10^5	2.9	3.0×10^4	~10,000
Round 2	600	1.6×10^4	16.4	2.6×10^2	~ 160

[†] difference to the mode of the signal distribution.

As in the previous experiment, the initial rates in cell lysate were measured after each round of enrichment (Figure 4.28). The data for the naive library was taken from the previous experiment. In the first round of enrichment, the proportion of clones with <10% activity compared to HG3.17 was reduced to 17%. In the second round this proportion increased slightly to 23% while the overall distribution did not change.

As before, no variants with activity 1.5 fold higher compared to the HG3.17 control were observed. Nonetheless, the six most active variants in lysate were selected for sequencing (Table 4.7). One was identical to the wild-type sequence of HG3.17 and two had synonymous mutations. Variant 2E10 had amino acid mutations Q143R, I150V and D179N. The locations of these mutations in the structure of HG3.17 is indicated in Figure 4.29. Q143 is a surface exposed residue and mutation to arginine introduces a charge which may improve the protein's solubility. D179 is only partially exposed while I150 is a buried residue mediating the interaction between two helices. Variants 3E10 and 3C12 both had the amino acid mutation A5T, but differed in a second amino acid mutation with Q77R and A181V respectively. Residue A5 is not observed in the crystal structure of HG3.17 and thus is unlikely to impact on the protein's structural integrity. Q77 is a surface exposed residue, and as in the case of Q143, mutation to arginine may have improved soluble expression of the protein. Finally, A181 is also located near the surface. Taken together, all the observed amino acid mutations were far from the active site and would be expected to have a small to neutral effect on the enzyme's activity.

The three variants which had amino acid mutations and HG3.17 were expressed, purified, and their Michaelis Menten parameters determined (Table 4.8). The catalytic efficiency $k_{\text{cat}}/K_{\text{m}}$ obtained for HG3.17 was in good agreement with the literature value (2.3×10^5 , [16]), although lower k_{cat} and K_{m} values were obtained respectively. As expected, the catalytic efficiencies obtained for the three mutant enzymes were close to wild type activity but

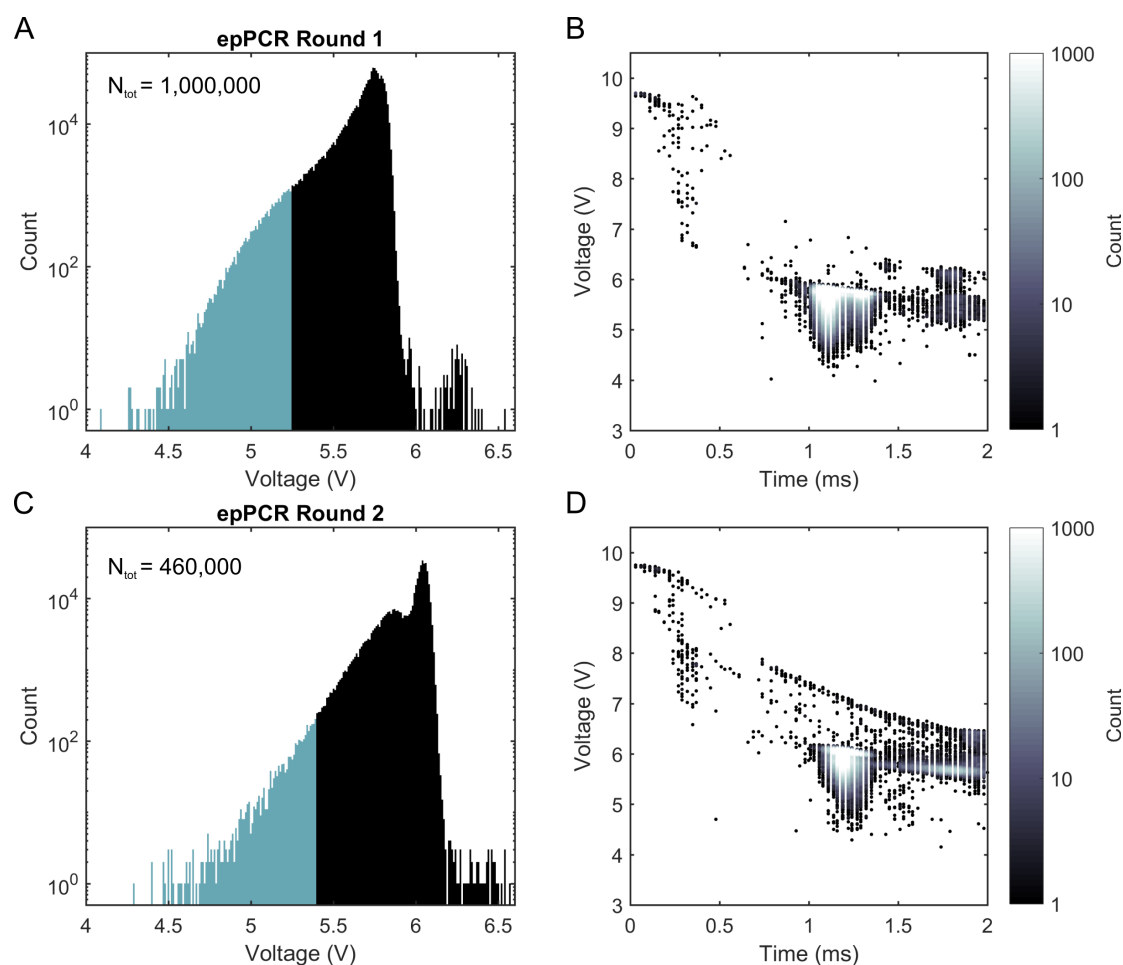


Fig. 4.27 Two rounds of AADS of the epPCR library of HG3.17. The first sort was performed at 1 mM substrate, the second at 0.75 mM in order to increase selection pressure. Droplets in the blue range were collected.

Table 4.7 Mutations found in the six most active variants in the cell lysate re-screening assay.

Variant	Mutations
2E10	T123C, A428G (Q143R), A448G (I150V), G535A (D179N)
3C12	G13A and A15T (A5T), C542T (A181V)
3E10	G13A (A5T), A229G (Q77R), T906C
4D09	A144G
4D11	
4D12	C474A

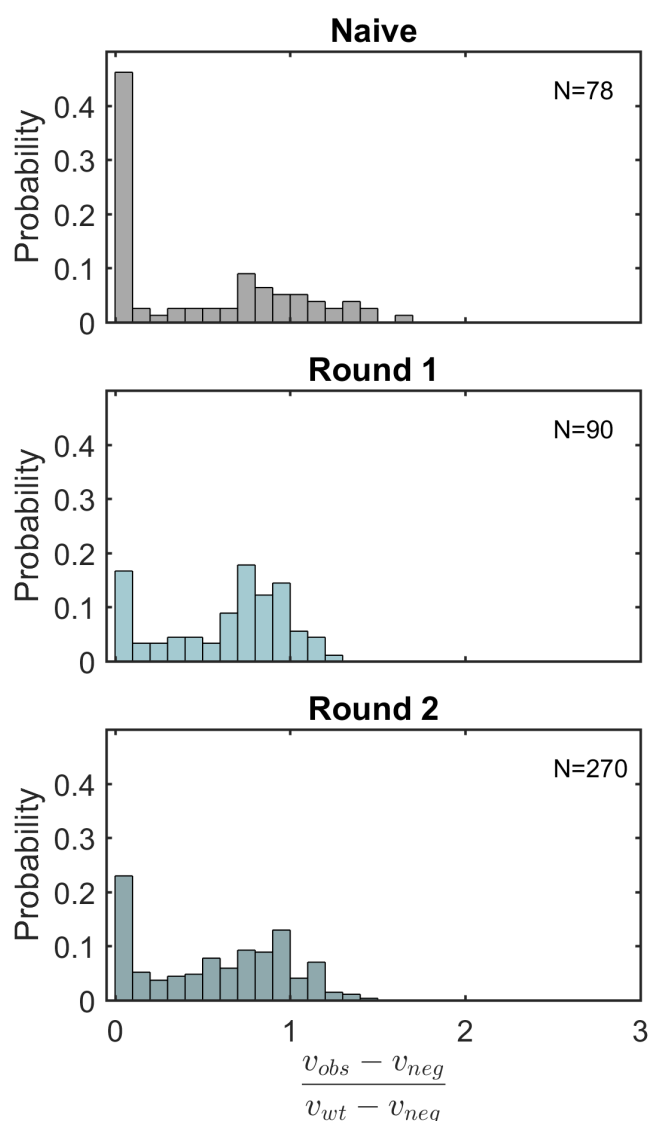


Fig. 4.28 Initial rates of clones from the epPCR library before and after each round of sorting. The rates are normalised using the average rates of three HG3.17 (v_{wt}) and three N20 (v_{neg}) samples respectively. As can be seen, the proportion of inactive clones is reduced after just one round of sorting.

about 2 fold lower. That these catalytic parameters were lower than for HG3.17 indicates that they were more active in lysate due to compensatory improved expression.

As a crude measure of soluble protein expression the size and intensity of the respective protein bands on gel were quantified in the insoluble and soluble fractions after overnight expression under the same conditions as for the droplet screening and cell lysis (see Figure 4.30). The sample loading of in particular of the pellet varied between samples and was therefore normalised relative to an endogenous *E. coli* protein band (indicated in the figure)

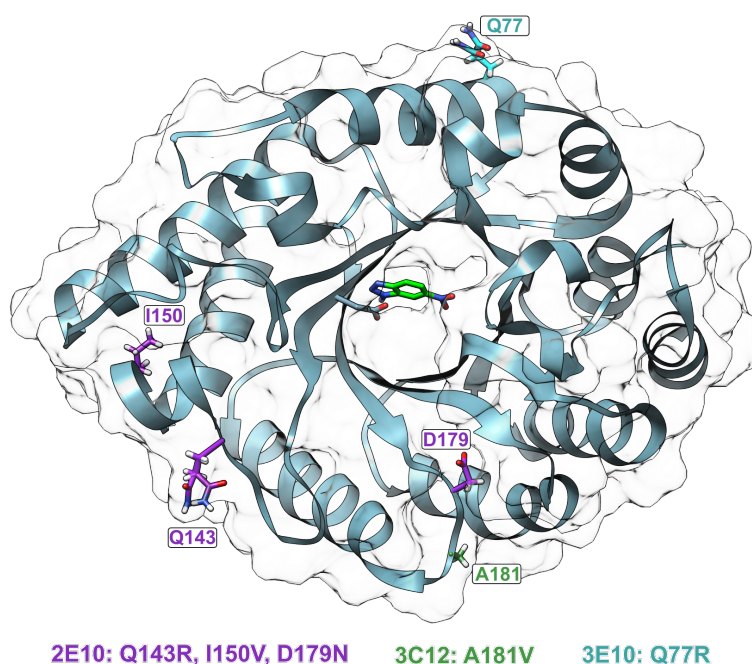


Fig. 4.29 Locations of the mutated residues in the three characterised variants within the structure of HG3.17 (PDB: 4BS0, complexed with transition state analogue 6-nitrobenzotriazole shown in green, [16]). The enzyme has a TIM barrel fold, with the catalytic aspartate located in the centre of the β -barrel H-bonded to the transition state analogue shown in green. The annotated residues were mutated in the characterised variants and are located far from the active site near the protein surface.

Table 4.8 Michaelis-Menten kinetic parameters of the variants selected in the cell lysate assay which had amino-acid mutations for substrate **2a**.

Enzyme	k_{cat} (s^{-1})	K_{m} (mM)	$k_{\text{cat}}/K_{\text{m}}$ ($\text{Ms})^{-1}$
HG3.17	243 ± 21	1.0 ± 0.2	$2.4 \pm 0.2 \times 10^5$
2E10	276 ± 61	2.6 ± 0.8	$1.1 \pm 0.2 \times 10^5$
3C12	163 ± 26	1.2 ± 0.4	$1.4 \pm 0.2 \times 10^5$
3E10	237 ± 66	1.7 ± 0.8	$1.4 \pm 0.4 \times 10^5$

Assay conditions: Enzymes at 5 to 6 nM, substrate 40 μM to 2 mM, 20 mM TrisHCl pH 7, 50 mM NaCl, 10% v/v MeOH.

to account for the difference. While soluble expression of HG3.17 started high at 88%, the analysis indicated 1.6 to 1.8 fold overall improved total soluble protein expression for variants 2E10, 3E10, and 3C12.

Improvements in soluble protein expression have previously been observed in directed evolution campaigns. For example, Aharoni *et al.* found that the initial rounds of evolution of the mammalian paraoxonase PON1 yielded variants that were 20 fold more active in lysate compared to the starting point, but had no improvements in its specific activity [221]. A similar trend was observed by Gielen *et al.* who evolved phenylalanine dehydrogenase in the only other directed evolution campaign that used AADS for screening [61]. The best variant isolated after two rounds of evolution showed 4.6 fold increased activity in lysate, but showed no change in k_{cat}/K_m but about 2 fold improved soluble expression. Similar to these, the selection criterion in this assay was 1) activity in single-cell lysates in droplets and 2) in bulk lysates in 96-well plates. Therefore, enrichment was determined by both the enzyme's catalytic properties as well as the concentration of functional enzyme, influenced by its stability under the conditions in both screening steps. Mutations that improve one at the cost of the other (a form of compensatory epistasis) will pass the selection threshold [222].

It is possible that there are additional stability requirement in droplets different from plates (*e.g.* interactions with a fluorous interface at the droplet boundary and fast turbulent mixing as droplets move through microfluidic channels). Furthermore, in droplets cell-to-cell expression variation will play a role (high variation reducing the likelihood of collecting the variant), whereas in bulk measurements cell-to-cell variation is masked as long as the population's mean expression improves.

In conclusion, a 100×larger epPCR library of HG3.17 was screened in a single round of screening than previously possible, but no catalytically improved variants were discovered. Given that HG3.17 was the endpoint of a previous directed evolution campaign, mutations to further improve catalytic activity may be very rare. At the same time, microfluidic droplets may place new stability constraints on the enzyme which require initial compensatory mutations, after which a new evolutionary trajectory becomes accessible in droplets. Importantly, it was shown that the droplet assay was efficient at enriching the proportion of active variants in the library. It was thus hypothesised that libraries with a bigger impact on the protein structure (a larger proportion of inactive variants) could be screened using this method to purge the least active variants. Therefore, the InDel libraries were constructed and screened, as described in the following.

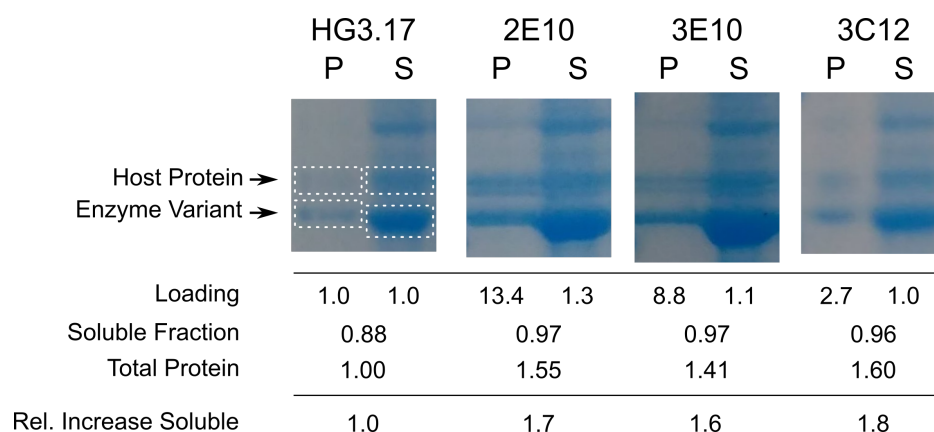


Fig. 4.30 Analysis of bulk soluble expression of HG3.17 and the characterised variants. Each protein band was assigned a value by multiplying the unadjusted average gray-scale intensity and area of the band (determined in ImageJ). The sample loading for each pellet and soluble lane was normalised to the respective lanes in the HG3.17 lane to determine the changes in soluble fraction and total protein expression.

Directed evolution using InDel libraries

As explained in Section 4.5, insertions and deletions have a larger impact on protein structure than substitutions and InDel libraries have thus a bigger proportion of inactive variants. Here, I explored if purging inactive variants at high-throughput is useful to identify positions that can tolerate InDels.

Generation of InDel libraries of HG3.17 using TRIAD In the deletion library, 3 consecutive base-pairs were deleted at a time. Therefore, the theoretical library size is equal to the number of base-pair triplets that can be deleted. Excluding the start-codon and 6xHis-tag that is 916 ($\sim 10^3$) in the case of HG3.17. While a library of 10^3 could be covered 1-2 times in 96-well plates, achieving high coverage is only practically possible using droplet microfluidics. In the case of the insertion library, an NNN triple base-pair was inserted. The library size in this case is $916 \cdot 64 = 58,624 \approx 6 \times 10^4$.

After library construction, ten colonies were picked and sequenced for the deletion and insertion libraries, respectively (see Figure 4.31). In the deletion library, all ten variants showed a 3 bp deletion and in two cases this led to mutation of one adjacent amino acid (due to out-of-frame deletion, see Figure 4.19). In the insertion library, six of ten variants had a 3 bp insertion, three variants had less than 3 bp inserted causing a frame-shift and one sequence was wild-type. This is a typical result of the cloning method used (oral communication Dr Stephane Emond). Assuming that only 60% of the obtained library had 3 bp

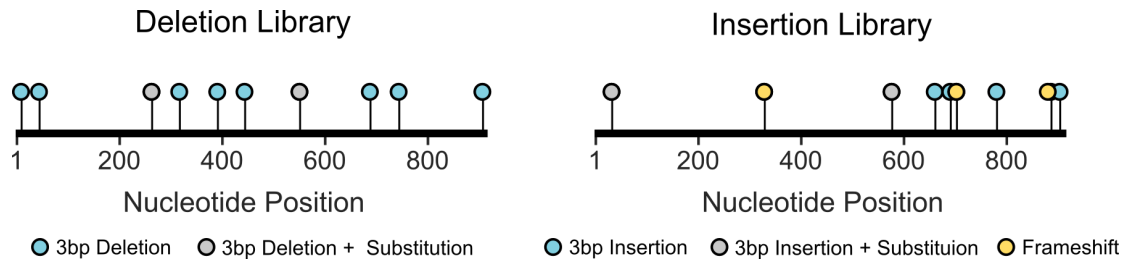


Fig. 4.31 The circles indicate the location of deletions and insertions, respectively. In the insertion library there was a slight bias towards insertions in the second half of the gene.

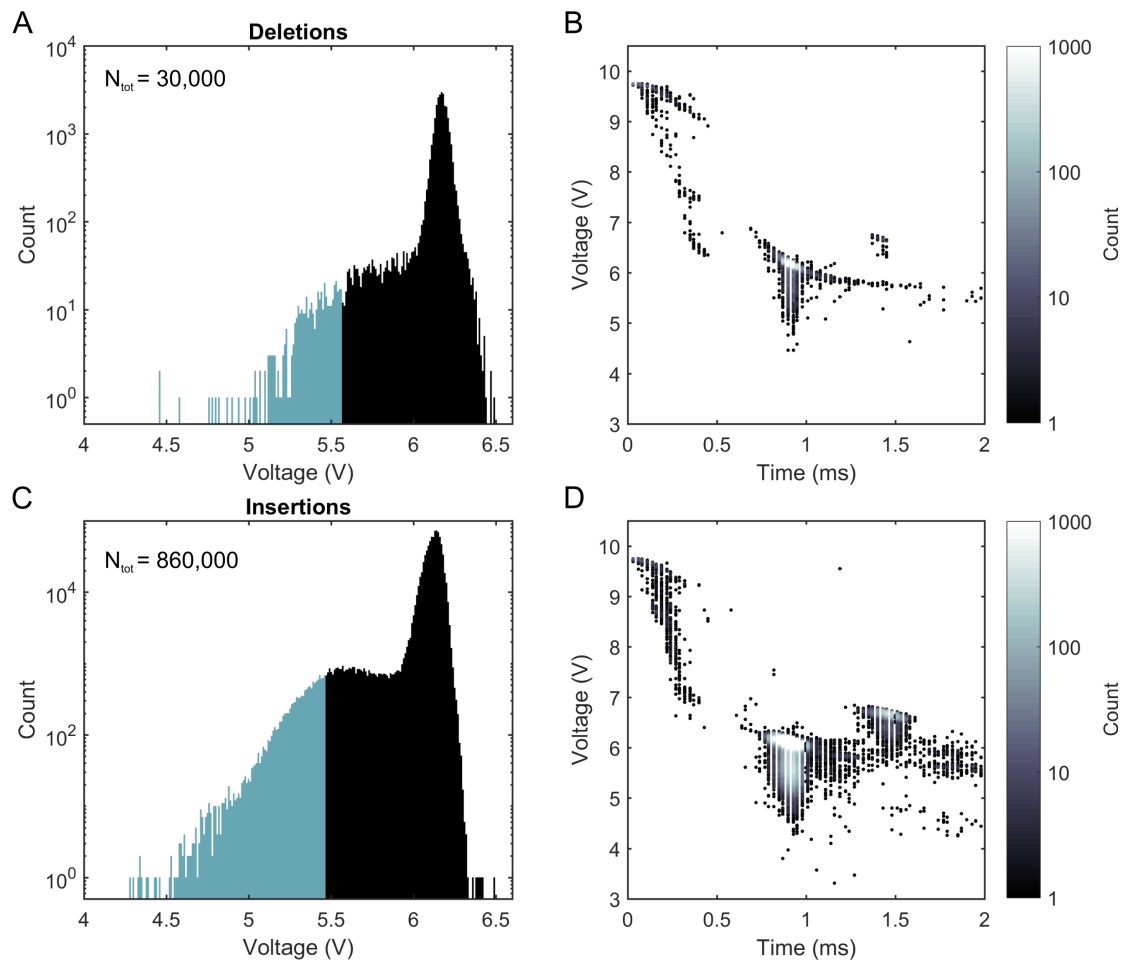


Fig. 4.32 AADS histograms for the deletion and insertion libraries. The shape of the histograms is distinctly different from the epPCR libraries, indicating a lower proportion of variants with detectable activity. Droplets in the blue range were collected (A and C).

insertions, the theoretical library size was still covered 30 fold in the final transformation step.

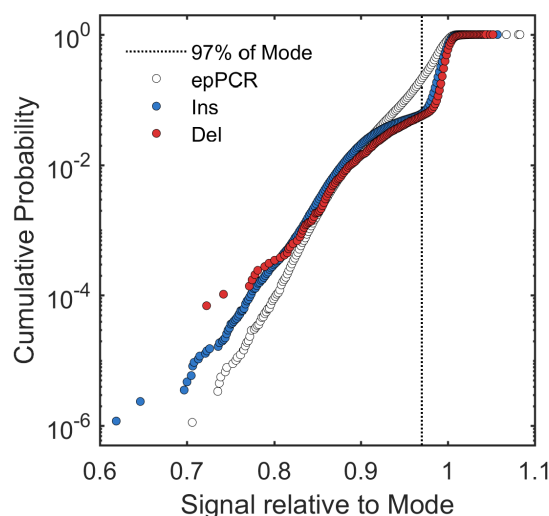


Fig. 4.33 The empirical cumulative probability function of the droplet sort for the epPCR, insertion and deletion libraries. As expected, the proportion of droplets with a signal below 98% of the signal distribution mode is reduced in the latter two, indicating that there are fewer variants with detectable activity.

Screening of InDel libraries Each of the libraries was transformed into electro-competent BL21-Gold(DE3) for protein expression and screened in droplets. The sorting histograms are shown in Figure 4.32. Using these histograms, the empirical cumulative probability function for each sort was calculated and compared to the first round of sorting of the larger (second) epPCR campaign. Figure 4.33 shows that the number of droplets with a signal below the mode of the distribution dropped off more quickly in both InDel libraries compared to the epPCR library. This suggests the fraction of droplets that contained variants with detectable activity was markedly reduced, as would be expected due to the more disruptive nature of InDels compared to point substitutions. At 97% of the distribution's mode there was a distinct “kink”, most likely corresponding to a transition between droplets with non-detectable and detectable activity. The probability to be below this threshold was 16.2% in the epPCR library but only 5.2% and 5.6% in the deletion and insertion libraries, respectively. Taking into account the Poisson distribution, this translates into 46%, 15% and 16% of the single-cells having detectable activity in droplets. Interestingly, at low signal (high activity), the InDel libraries had slightly increased probabilities, suggesting there may be more active but rare variants present compared to the epPCR library.

In the case of the deletion library, 3×10^4 droplets were screened allowing over 10 fold coverage of the theoretical library size in just 10 min of sorting (Table 4.9). In the case of the

insertion library, 8.6×10^5 droplets were sorted over ~ 2.5 h covering the theoretical library size 5 fold (3 fold if discounting for the frame-shifted variants).

Table 4.9 Number of variants analysed during each round of enrichment of the epPCR library of Kemp eliminase HG3.17.

Sample	Threshold [†] (mV)	Cells Analysed $N_{\text{tot}} * \lambda$	Coverage C	Droplets Collected N_{coll}	Output Library N_{out}
Deletions	590	1.0×10^4	11.5	4.0×10^2	~ 40
Insertions	690	3.0×10^5	5.1	1.2×10^4	~ 400

[†] difference to the mode of the signal distribution.

Individual variants were tested in cell lysate assays and the resulting histograms are shown in Figure 4.34. Prior to enrichment, 90% of the deletion library had less than 10% activity in cell lysate compared to HG3.17 wild-type activity. This is in good agreement with the prediction from the cumulative probability function of the sorting histogram⁵. The proportion of low-activity (and inactive) variants was reduced to 40% after droplet sorting suggesting effective enrichment. Unfortunately, too few colonies from the naive insertion library prior to sorting re-grew to perform the cell lysate assays. However, given that the cumulative probability functions of the insertion and deletion libraries were very similar, it is likely that the proportion of low-activity variants was similar. However, the enrichment of activity was less effective for the insertion library with low-activity variants retained at 60%.

As previously, the most active variants in cell lysate were sequenced. In particular the deletion library had a number of promising candidates, at 2 fold and higher activity compared to wild-type. Yet, all of them were wild-type except for variant 4A09, which was 2.7 fold more active in lysate. This variant had a deletion of A181 (DelA181) with adjacent mutation S182G. Interestingly, variant 3C12 from the epPCR library carried a mutation in the same position, A181V. Variant 4A09 was purified and its Michaelis-Menten parameters determined. Its $k_{\text{cat}}/K_{\text{m}}$ was 10 fold reduced compared to HG3.17 compensated for by an estimated 4.5 fold improved soluble expression (Figure 4.35) and possibly other stabilising factors. This is particularly interesting, as discussed earlier improved expression or at least adaptation to droplet conditions is likely necessary, while the deletion itself together with starting at a reduced initial activity may open unprecedented mutational trajectories for improved catalytic activity, further explored below.

⁵Which was 15%, $N = 20$ in the cell lysate was small. If one more variant would be tested and found to be active, the cell lysate proportion would be 14%.

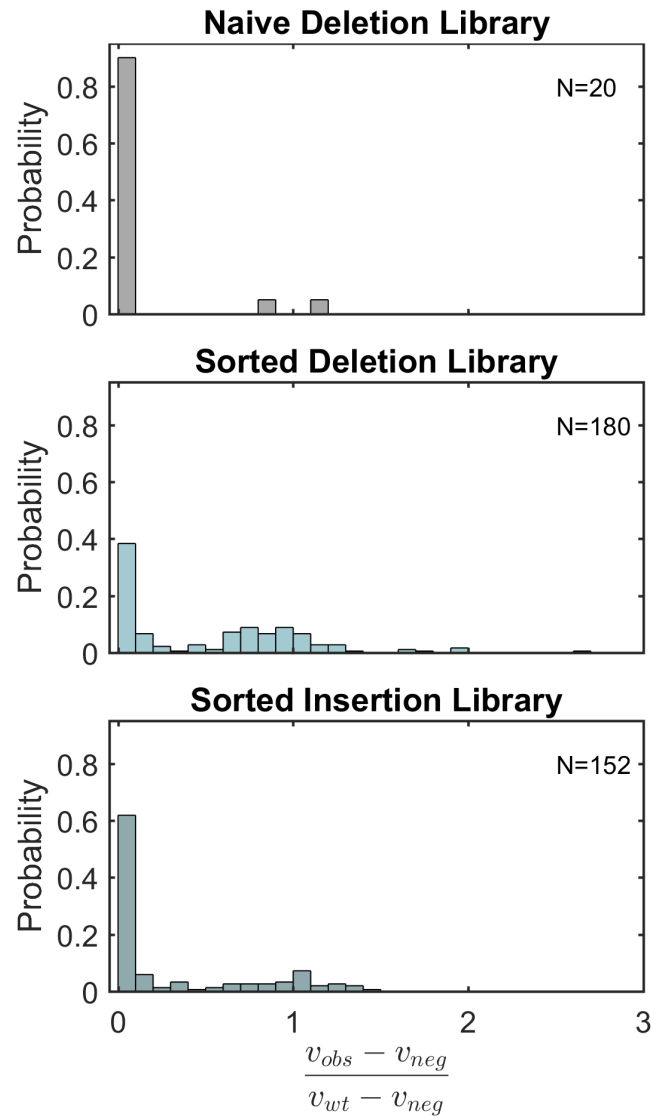


Fig. 4.34 Initial rates in cell lysates normalised using the average rates of three HG3.17 (v_{wt}) and three N20 (v_{neg}) samples, respectively. As can be seen, the proportion of inactive variants is markedly reduced after sorting the deletion library. Too few variants of the naive insertion library re-grew to assess the enrichment, but assuming a similar distribution prior to sorting, it was less efficient than for the deletion library.

Table 4.10 Michaelis-Menten kinetic parameters of deletion variant 4A09 (DelA181, S182G) for substrate **2a**.

Enzyme	k_{cat} (s^{-1})	K_{m} (mM)	$k_{\text{cat}}/K_{\text{m}}$ ($\text{Ms})^{-1}$
HG3.17	243 ± 21	1.0 ± 0.2	$2.4 \pm 0.2 \times 10^5$
4A09 (DelA181, S182G)	63 ± 20	3 ± 1	$2 \pm 1 \times 10^4$

Assay conditions: Enzyme 4A09 at 10 nM, substrate 40 μM to 2 mM, 20 mM TrisHCl pH 7, 50 mM NaCl, 10% v/v MeOH.

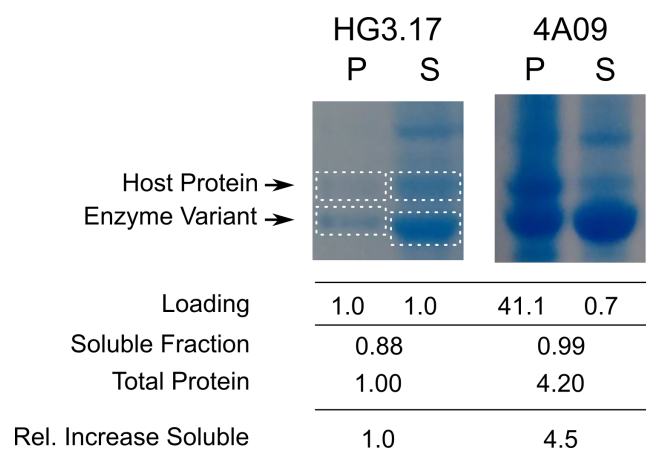


Fig. 4.35 Analysis of relative protein expression according to the same method described for Figure 4.30.

The identification of neutral InDels may open unprecedented mutational trajectories

The majority of variants sequenced from all library sorts so far were wild-type. Given that the starting activity of HG3.17 was high and most mutations detrimental, this was not wholly unexpected. In fact, it could be seen as a quality marker of the droplet screen to enrich wild-type sequence present in the libraries. To gain a more comprehensive understanding of the outcomes of the InDel library screening, two entire 96-well plates were sequenced, one for the deletion and one for insertion library, both post-sorting.

It became apparent that the majority of variants on the deletion plate were wild-type HG3.17 (45 of 83), see Figure 4.36. No wild-type had been found in the ten randomly sequenced variants from the naive library. The next most frequent variants (13) had deletions in positions 305-307 (C-terminus) retaining an average activity of $93 \pm 22\%$ in lysate relative to HG3.17. These C-terminal residues do not appear in the crystal structure of HG3.17, *i.e.* they are unstructured and unlikely to have a large impact on the protein structure. DelA66 was found 8 times, with an average of $13 \pm 3\%$ lysate activity relative to HG3.17. Deletion of A66 without adjacent mutations is possible through removal of 196-GCG-198 (in-frame) or 197-CGG-199 (out-of-frame), the resulting sequences being indistinguishable. Assuming no bias in the library, the expected frequency for two independent triplet base deletion is $2/916$; *i.e.* DelA66 was enriched about 44 fold. Interestingly, DelP233 was found twice (10 and 15% activity) along with a DelP233, Q234E variant (only 2% activity). Proline is rigid (φ is locked to -65° , because its side chain is attached to the α -amine) and often observed at the end of secondary structure elements when the back-bone reverses its direction (*e.g.* in β -turns). Its introduction or removal would therefore generally be expected to be disruptive, but this depends on the structural context. DelS182 was showed 107% activity in cell lysate, further corroborating that positions 181-182 are tolerant towards both deletions and substitutions. Furthermore, two variants with substitutions were found, which could have arisen from errors during the cloning procedure.

About 23 wells of the insertion plate did not yield sequences, which may have been due to a contamination and explain in part why the enrichment appeared to be less efficient compared to the deletion plate. Again, a large proportion of the sequences (20 of 67) were wild-type. In this library, one of the ten randomly sequenced variants had been wild-type, indicating an approximate 3 fold enrichment after droplet sorting. Again, an insertion into the C-terminal region was found (Ins305L, 103%), but more strikingly Ins61F (27%), Ins67L (18%) and Ins68G (60%) were identified, indicating that the region 61-68 is tolerant towards both insertions and deletions. Another region of interest emerged at positions 229-233: Ins230G, A230S (13%) and S229I (38%). Ins106Q (13%) was identified along with 2x Ins106R, L106V (1%) and another substitution A181V (109%) was found once more.

Sequencing Result Deletion Plate Post-Sorting

WT	ΔQ305	WT	WT	WT	WT	WT	ΔQ222	ΔQ305	WT	ΔA214	ΔP233
ΔA66		Silent	ΔQ222		WT	HG3.17	WT	WT	ΔQ305	No Insert	ΔS182
WT	WT	WT	ΔY296	WT	WT		ΔS307	ΔG306	WT	ΔP233	WT
ΔS307	A218V	ΔA66	WT	WT	ΔQ222		WT	ΔA66	ΔS307	WT	
ΔQ305	WT	V109M	WT	WT	WT	N20	WT	ΔP233, Q234E	ΔA66		WT
WT	ΔA199	ΔA66	ΔA66	No Insert	ΔQ305		WT	ΔA66	WT	WT	1NT Del
WT	ΔA230	WT	WT	WT	Q305H, ΔG306		WT	WT	WT	ΔA66	ΔS307
WT	WT	ΔS307		WT	WT	WT	WT		ΔW128		ΔS307

Normalised Activity in Cell-Lysates

0.55	1.13	0.94	0.75	1.37	0.54	0.90	0.00	0.89	0.82	0.00	0.10
0.08		0.92	0.01		1.79	0.93	0.78	1.17	0.50	0.00	1.07
0.88	1.02	0.83	0.00	1.22	1.02	1.33	1.16	0.69	1.12	0.15	1.15
1.01	0.70	0.10	0.94	1.42	0.01	1.09	0.89	0.13	1.13	0.97	
0.57	0.85	0.35	1.09	1.41	1.43	0.00	1.13	0.02	0.16		1.17
0.82	0.10	0.12	0.14	0.00	0.94	0.00	1.19	0.15	0.98	1.03	0.00
0.78	0.01	0.89	0.79	1.11	0.93	0.00	1.13	1.01	0.80	0.13	1.24
0.74	0.74	0.93		0.52	1.25	1.04	0.67		0.00		1.00

Fig. 4.36 One 96-well plate of the deletion library post-sorting was sent for sequencing. Wild-type HG3.17 (WT) was the most commonly found sequence (54%). Black squares indicate no growth or a contamination. Normalised activities (relative to HG3.17 and N20 controls) in cell lysate are shown in the bottom panel.

Sequencing Result Insertion Plate Post-Sorting

WT	Ins61*		Ins205P	Ins257G		WT	Ins108G	1ntDEL	W128C, Ins129R		Ins298G
Ins143S	WT		Ins106R, L106V	WT		WT Control	2ntINS			2ntINS	
2ntINS		WT	Ins205P		Ins305L		Ins71T	A181V	Ins195P, Q195E	1ntDEL	
WT	Ins288S, F287L	WT	2ntINS	Ins71T	WT		WT	W128C, Ins129R			Ins227D
WT		2ntINS		Ins106Q		N20 Control		Ins229S	Ins196A	Ins87T	
WT	Ins106R, L106V	Ins71L		Ins205P	Ins121L			2ntINS	Ins196K	WT	Ins211*
Ins210*	S229I		WT	Ins96S	Ins67L		WT	Ins71T	WT	2ntINS	Ins61F
WT	Ins230G, A230S		WT	Ins148V	Ins68G	2ntINS	WT		WT	2ntINS	WT

Normalised Activity in Cell-Lysates

0.90	0.21		0.00	0.00		1.12	0.00	0.00	0.00		0.00
0.11	0.76		0.01	0.00		1.31	0.00			0.00	
0.00		1.04	0.00		1.03	0.98	0.00	1.09	0.00	0.00	
1.00	0.00	1.07	-0.01	0.00	1.31	0.83	0.59	0.00			0.00
1.06		0.00		0.13		0.00		0.00	0.09	0.00	
1.08	0.01	0.00		0.00	0.00	0.00		0.00	0.31	1.13	0.00
0.00	0.38		1.04	0.00	0.18	0.00	1.05	0.00	1.18	0.00	0.27
0.80	0.13		1.01	0.00	0.60	0.00	0.64		1.18	-0.01	0.92

Fig. 4.37 One 96-well plate of the insertion library post-sorting was sent for sequencing. Wild-type HG3.17 (WT) was the most commonly found sequence (30%). Black squares indicate no growth or a contamination. Normalised activities (relative to HG3.17 and N20 controls) in cell lysate are shown in the bottom panel.

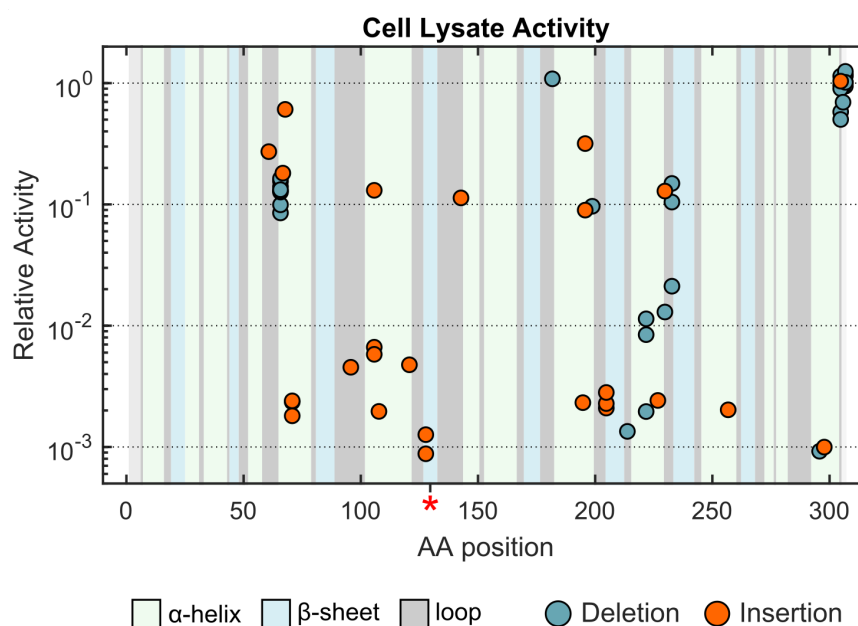


Fig. 4.38 Relative cell lysate activity of InDels versus amino acid position and secondary structure elements in HG3.17. As can be seen, all InDels that have near neutral to small effect on activity are located at the start or end of secondary structure elements. Some InDels are tolerated in the α -helices but reduce activity 100 fold.

Table 4.11 summarises the variants with mutations in nearby locations which were found multiple times or in more than one library. The libraries were all constructed independently and based on different cloning strategies, thus finding mutations in similar locations is a strong indication that these variants were selected for during the droplet sorting campaigns. The location of the mutations in the secondary and tertiary structure of HG3.17 are indicated in Figures 4.38 and 4.39, respectively.

Enrichment of the C-terminal mutations was also an indication of positive selection. Q305 is the first C-terminal residue that does not appear in the crystal structure (no electron density recorded) and S307 is the last residue at which deletion or insertion can occur without affecting the restriction site used to clone the gene into the vector used for screening. Thus, it is unlikely that changes in these three residues affect the proteins structure as is shown by the near-wild-type activity levels in lysate.

The overall structure of HG3.17 is a TIM-barrel, *i.e.* there is a core of eight parallel β -sheets connected by intervening α -helices (Figures 4.38). Regions 61-68, 181-182, 196-199 are located at the beginning or end of an α -helix whereas residues 229-233 comprise the loop connecting α -helix 6 to β -sheet 7. No InDels with residual activity were identified in the β -sheet core. This is consistent with other studies of InDels. Kim and Guo found in a

Table 4.11 Summary of mutations that were found more than once and in different libraries.

Position	Library			Total
	epPCR	Del	Ins	
61-68		DelA66 (8x)	Ins61F Ins67L Ins68G	11
106-109		V109M	Ins106Q Ins106R, L106V (2x)	4
143	Q143R		Ins143S	2
181-182	A181V	DelA181, S182G DelS182	A181V	4
196-199		DelA199	Ins196A Ins196K	3
229-233		DelP233 (2x) DelP233, Q234E	S229I Ins230G, A230S	5
305-307		DelQ305 (5x) Q305H, DelG306 DelG306 DelS307 (6x)	Ins305L	14

systematic analysis of 200 protein structures that the majority of InDel residues were exposed to solvent, *i.e.* near the protein surface and that a third of InDels were located in disordered regions [223]. Gonzalez *et al.* studied single amino acid insertions and deletions in TEM-1 β -lactamase [224]. They found that insertions and deletions were tolerated in loops and at the ends of α -helices and β -sheets (10 to 100 fold drop in fitness) but not within them (> 100 fold drop in fitness). Similar locations and changes in fitness expressed as activity in cell lysate were observed for HG3.17. The locations of the mutations can be further rationalised by looking at the B-factors, which are an indicator of flexibility⁶. These indicate that the surrounding α -helices are more flexible than the β -sheet core, which is where most mutations were found. The exception is α -helix 2, which appears to be more rigid.

In light of these considerations, regions 61-68 and 106-109 seem of particular interest. As can be seen in Figure 4.40A, L106 and V109 both mediate interactions with the loop connecting α -helix and β -sheet 3 (dark blue) and a short α -helix in the loop connecting α -helix and β -sheet 2 (purple). At the N-terminus of this short helix, residue Q52 protrudes into the active site forming a hydrogen bond with transition state analogue 6-nitrobenzotriazole. This interaction was suggested to facilitate catalysis by stabilising the negative charge developing on the oxygen ([16]). In the original computational design of this enzyme (HG2), lysine was placed at this position intended to bind the nitro-group of the substrate [15]. However, the crystal structure of HG2 revealed that the transition state analogue was flipped with respect to the design. During the directed evolution towards HG3.17, the lysine the first mutated to histidine (K52H) and further to glutamine (H52Q). A Q52A mutant of HG3.17 was 50 fold less active compared to wild-type HG3.17, establishing the importance of the observed interaction. Deletions and insertions at positions G61, A66, G67, and A68 would also affect interactions with the Q52 α -helix as well as the loop connecting to β -sheet 2 (cornflower blue). Together both loops adjacent to the Q52 α -helix also mediate access to the active site as they partially cover it (4.40B). Changes in the dynamics of these loops will thus affect substrate entry and product release.

Therefore, a region of interest capable of accepting insertions and deletions has been identified. One beneficial mutation, T107I was observed in the 106-109 region previously by Blomberg *et al.*, but none in the region 61-68. While the introduction of InDels reduced the activity in cell lysate, they may open up new evolutionary trajectories. A study of InDels in evolving proteins of *Drosophila* showed that incorporation of InDels triggers new substitutions with a high fraction of them occurring in previously conserved positions [227]. It may be interesting to construct combinatorial libraries of the mutations in regions 61-68 and 106-109 and variant 4A09 (Del181A, S182G), which was found to have improved soluble

⁶The B-factor measures the attenuation of x-ray scattering due to thermal motion of the atoms [225, 226].

expression. A better expressing InDel variant could then be subjected epPCR to introduce substitutions and explore if new sequence space has become accessible to improve the catalytic efficiency beyond the levels of HG3.17.

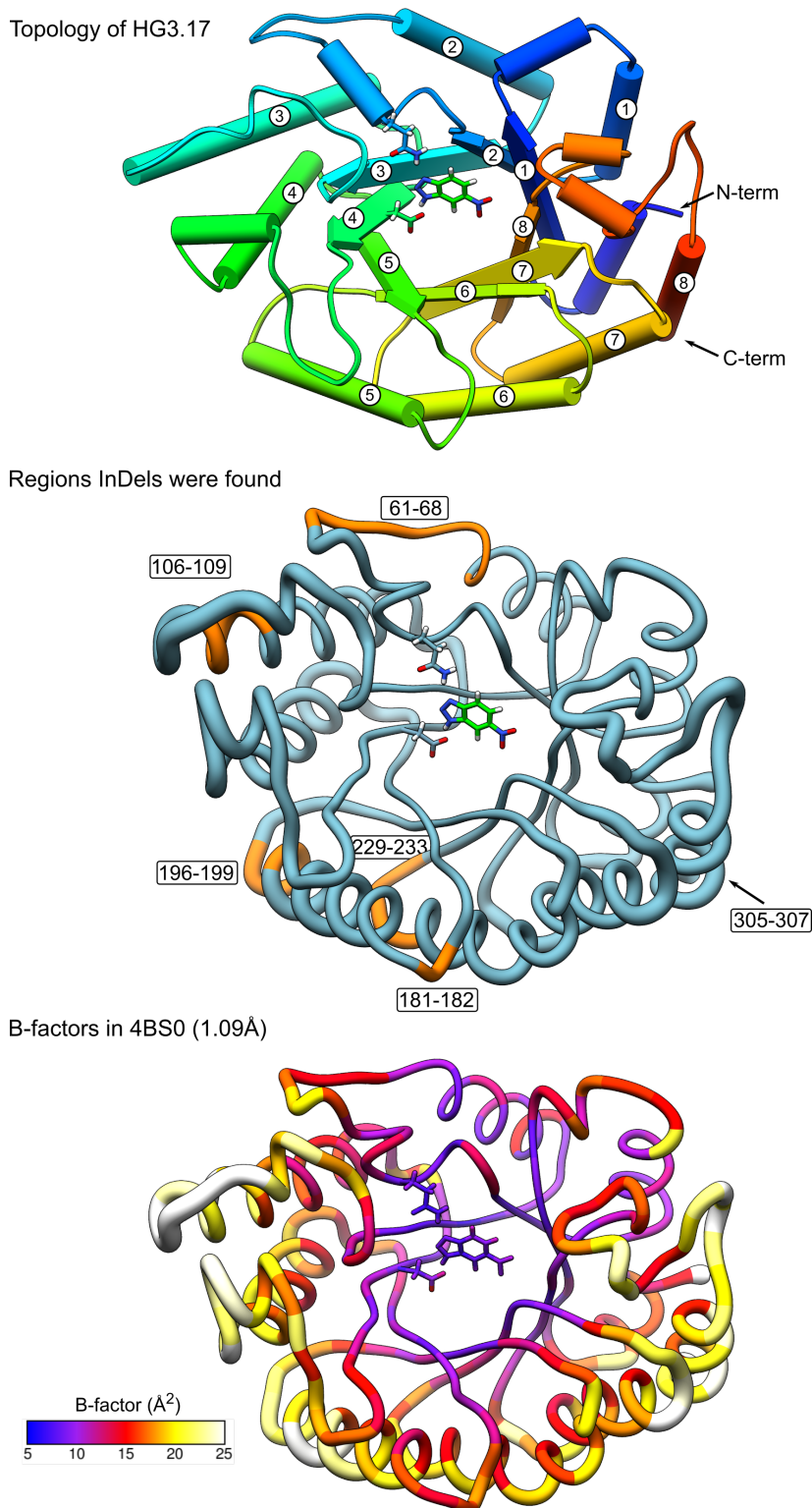


Fig. 4.39 Topology of HG3.17 (top panel) and location of regions identified to accept InDels (orange, middle panel). The thickness of the backbone in the middle and bottom panel is indicative of the average B-factor of the residue at the given position. The bottom panel is additionally coloured according to the B-factor at the respective position. PDB: 4BS0 ([16]).

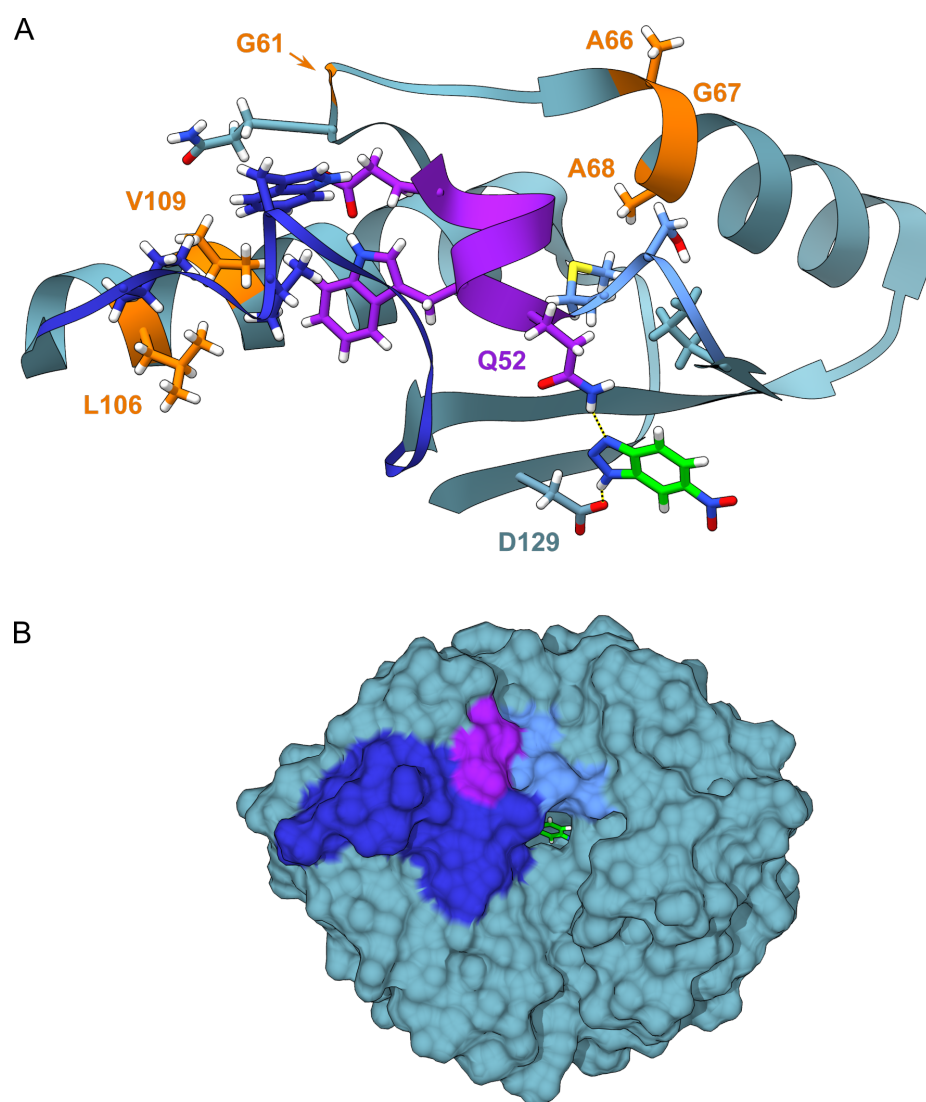


Fig. 4.40 A: The mutations found during the InDel screening campaign (orange residues) are likely to affect interactions with two loops (dark blue, cornflower blue) which mediate access to the active site and, crucially, a short α -helix (purple) from which residue Q52 protrudes. This residue forms a hydrogen bond with the transition state analogue 6-nitrobenzotriazole. B: Top view of HG3.17 showing how the indicated loops and α -helix restrict access to the active site. PDB: 4BS0 ([16])

4.6 Conclusions

In order to screen for Kemp eliminase activity, a droplet-based microfluidic assay was developed. The assay was based on the substrate 5-nitro-1,2-benzisoxazole (**2a**). The relative ease of the detection of its reaction product has made **2a** the choice substrate amongst 1,2-benzisoxazoles. In contrast to other 2-hydroxybenzonitriles, 2-hydroxy-5-nitrobenzonitrile has a strong absorption band in the near UV range at 380 nm [94]. For this reason, **2a** was chosen as the substrate to test in droplets. The droplets used were 70 μm in diameter (180 pL in volume). It was found that the reaction product exchanged between droplets on a timescale of several hours, which limited the ability to distinguish droplets containing active or no enzyme after long incubation times.

In spite of this limitation, enzymatic activity of the Kemp eliminase HG3.17 was observed from single-cell lysates in droplets. Two enrichment experiments of HG3.17 over a negative control, the esterase N20, were successful with efficiencies of 97% and 99% compared to the maximal possible enrichment. This is the first report on detecting Kemp eliminase activity in microfluidic droplets. Screening of the metagenomic library SCV did not yield any events with absorbance above background. Therefore, given the product leakage, the assay was not sensitive enough for functional metagenomic screening.

Since HG3.17 was shown to enrich over a negative control, it presented itself as an interesting case to screen mutant libraries using this assay. Blomberg *et al.* evolved HG3.17 from HG3 over 17 rounds and it remains the best Kemp eliminase reported to date [15, 16]. The 96-well plate assay used to perform this evolution had a throughput of only 10^3 mutants per round. The droplet assay established outstrips this by two orders of magnitudes (based on sorting for 60 min at 100 Hz). A higher-throughput was previously shown to overcome limitations in directed evolution as in the case for another *de novo* enzyme by Obexer *et al.* [47, 79]. Therefore, three libraries of HG3.17 were prepared. One using epPCR to obtain point mutations resulting in amino-acid substitutions and two libraries deleting or inserting 3 base-pairs, respectively. These libraries contained all possible single amino acid insertions and deletions, as well as all possible substitutions of either the N-terminal or C-terminal neighbouring amino acid.

The epPCR library ($\sim 1.5 \times 10^5$) was screened twice using different strategies to fine-tune the stringency of droplet collection. The fraction of low-activity variants was consistently reduced post-sorting, indicating positive selection. It is interesting to note that the low activity fraction ($>10\%$ or no activity) was reduced from about 60% to 17-20% and remained at about this level even after two rounds of sorting. There are different sources of bias and error in droplet sorting, which can explain why low-activity variants are carried over. Here, the carry-over can be fully explained by co-encapsulation of more than one cell in droplets due

to the Poisson distribution (detailed explanation in Section 6). For example, when collecting 400 droplets from a droplet population with $\lambda = 0.35$ one can predict that close to 300 of these would contain only one (positive) cell, while the remainder would contain additional “passenger” cells corresponding to about 20% of all cells collected. Therefore, these sorting outcomes could be improved in future experiments by reducing λ – with the limitation that a low λ would require longer sorting times, which are limited for this assay. A promising strategy could be to perform a first round of sorting using $\lambda = 0.35$ and reduce it to *e.g.* 0.1 in a second round (8% carry-over predicted). Changing λ does not affect the stringency of the sort. This could be independently changed by an increased collection threshold or reduced substrate concentration as explored for these epPCR library sorts.

The epPCR sorting campaigns did not yielded variants of HG3.17 which were likely to have improved soluble expression, but none with improved catalytic parameters were found. An often invoked metaphor for the directed evolution of an enzyme is an up-hill walk in a fitness landscape [228]. Using this metaphor, it may be that HG3.17 was atop a local fitness peak, which could not easily be overcome at the mutational load (2.7 mutations/gene) used. In a future study, a higher mutational load combined with the high throughput may allow the “jumping” to a different fitness peak due to *e.g.* epistatic effects between several mutations.

Here, I chose to follow an alternate route and screen the much less studied insertion and deletion libraries. The insertion or deletion of an amino acid can be thought of as a leap in sequence space. In general this may be detrimental initially, but starting from a new point in sequence space, an adaptive up-hill walk may be able to follow a previously inaccessible trajectory. The InDel screening campaign yielded one variant with improved soluble expression and revealed five locations in HG3.17 which were able to accept InDels with a retention of at least $\geq 10\%$ activity in cell lysate. Two of these are particularly interesting, 61-68 and 106-109. They are in α -helices that mediate interactions with a short α -helix wedged between them that positions a glutamine residue in the active site of the enzyme, which forms a hydrogen-bond with the substrate. Ins68G in particular also displaces a second-shell alanine, which is in contact with a first-shell proline.

In future experiments one of the highlighted InDel variants combined with mutations that improved soluble expression could serve as a starting point for a point-substitution library. A general strategy emerging from this could be to alternate a few rounds of directed evolution with substitutions, which may better mimic natural evolution. As mentioned previously, InDels to substitutions occur at a ratio of 1:5 in the non-coding regions of bacterial genomes ([210]), which could be used as a guide on how often to alternate between the two types of mutations in an extended directed evolution campaign.

Chapter 5

A novel fluorogenic Kemp substrate to screen for Kemp eliminases in the metagenome using FADS

5.1 Abstract

The work presented here builds on the search for Kemp elimination substrates that do not exchange between droplets in droplet-microfluidic assays. I found that one substrate, 5-azido-1,2-benzisoxazole (**6a**), yields a fluorescent reaction product after conversion by Kemp eliminase HG3.17 and subsequent exposure to UV-light. The evidence presented here supports that the first reaction step yields the expected Kemp elimination product. While the origin of fluorescence requires further clarification, it was possible to enrich the Kemp eliminase HG3.17 over a negative control from single-cell lysates in 1.8 pL droplets using substrate **6a** and the FADS instrument developed in Chapter 2. As opposed to the absorbance-based assay discussed in Chapter 4, the fluorescence-based assay detected substrate turnover in functional metagenomic screening and the apparent hit-rate indicated that Kemp eliminase activity was common in the metagenome. After DNA recovery, a much larger number of colonies than expected was obtained and re-screening in well-plate format revealed that these were library variants containing deletions in the vector of up to 1 kbp. These are likely to have been enriched due to biological biases favouring the amplification of small vectors. Droplet screening using a subset of the SCV metagenomic library containing fewer vectors with deletions reduced the number of false positives and re-screening combined with sequencing yielded a potential hit, an ORF predicted to encode for a class IV adenylyl cyclase. This is promising and, considering the high apparent hit-rates observed, indicates that future screenings using

high-quality metagenomic libraries could yield many promiscuous enzymes catalysing the Kemp elimination.

Contributions: *I discovered the red fluorescence of the reaction product of compound 6a, investigated and characterised the reaction in both droplets and well-plates under different conditions using UV-VIS spectroscopy. Compound 6a and its precursors 5a and 6a were synthesised by Dr Josephin Holstein based on our discussions. She provided support by lyophilising the reaction products for submission to the analytical services at the Department of Chemistry (described in Section 7.3.5). The analysis and interpretation of the obtained data is my own. I optimised the droplet screening conditions, performed the enrichment and metagenomic screening experiments, recovered the DNA, re-screened the recovered variants in plates, analysed the sequencing data, modelled the structure of variant G06, as well as performed all of the analysis and interpretation of the data.*

5.2 A moment of serendipity: a fluorogenic Kemp substrate

The idea of azide **6a** (Figure 5.1) was to have the flexibility to add different functional groups *via* strain-promoted azide–alkyne cycloaddition (copper-free “click” chemistry) in dilute aqueous solution at room temperature, such as in compound **7a**. Both **6a** and **7a** were converted using purified HG3.17 and the resulting solutions used for a leakage test. However, due to small absorption coefficients the initial signal difference was too small to quantify leakage (data not shown). Surprisingly, the top layer of the droplets acquired a red colour two days from their generation (Figure 5.2A). The droplets were investigated by fluorescence microscopy and it was revealed that a red-emitting fluorophore had formed in both samples. In particular for substrate **6a**, it appeared to the eye that there were two equally sized populations of differing brightness, as would have been expected. Manual analysis of each 16 droplets from the brighter and darker population showed an unadjusted average brightness of 45 ± 5 and 35 ± 5 (of 255), respectively. Within the layer which was colourless to the eye only some of the droplets were fluorescent at all. Why there were two layers was not clear. Taken together, these initial serendipitous observations merited further investigation of the two substrates.

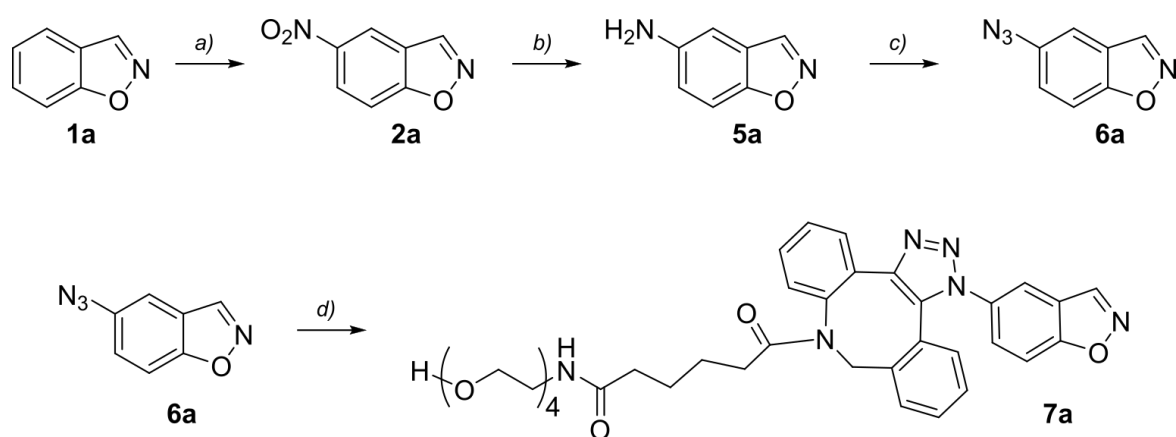


Fig. 5.1 Synthesis of **6a** and **7a**. Conditions: *a*) $\text{H}_2\text{SO}_4/\text{HNO}_3$ *b*) $\text{SnCl}_4/\text{SnCl}_2$ *c*) $\text{NaNO}_3/\text{NaN}_3$ *d*) 20 mM TrisHCl pH 7 and 50 mM NaCl. These reactions were performed by Dr Josephin Holstein.

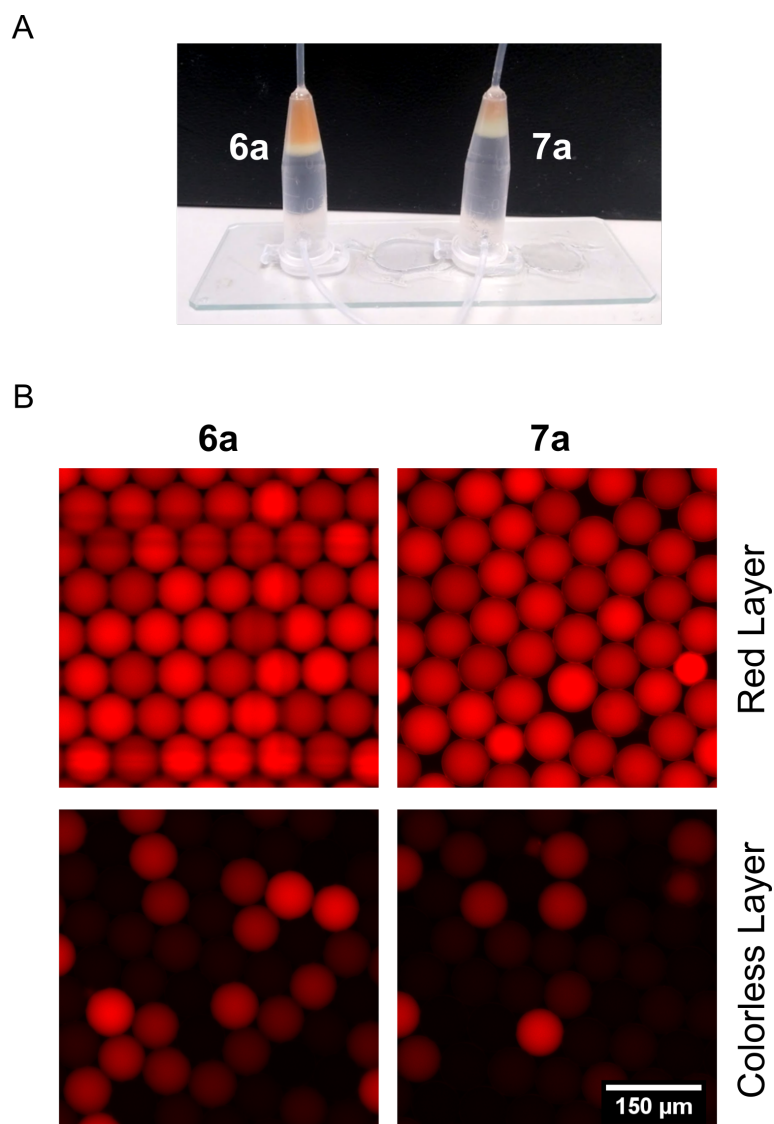


Fig. 5.2 Substrates **6a** and **7a** were converted under basic conditions and tested for leakage. *A*: The droplets generated for the leakage test turned red after two days at room temperature. In each sample two layers, one red and one colourless, were observed. *B*: Fluorescence microscopy revealed red fluorescence of the droplets in the red layer. In the colourless layer, droplets were not fluorescent at all or brightly fluorescent (rather than being of *e.g.* intermediate fluorescence). Excitation: 531/40 nm, emission: 593/40 nm.

5.2.1 Reproducing the initial observations

The first step in investigating the new fluorophore was to reproduce the initial observation. Two parallel experiments were set up: one in 384-well plate format and one in droplet format.

In the 384-well plate, substrates **6a** and **7a** were each incubated with the purified enzymes HG3.17, N20, and in buffer only. The reactions were monitored using fluorescence excitation and emission scans. A fluorescent signal emerged over the course of several days in all samples, but most strongly for HG3.17. The fluorophore had an excitation maximum at 500 nm and an emission maximum at 600 nm for both substrates. After 5 days of incubation, the emission at 600 nm was 100× higher in the HG3.17 sample compared to both N20 and buffer for substrate **6a**. For substrate **7a** the ratio was 50×.

Droplets of 70 µm diameter were generated for each substrate with and without HG3.17 using the same conditions as in the plate. The samples were stored separately in the dark to protect the fluorophore from potential bleaching. Surprisingly, after four days of incubation, none of the samples had developed red fluorescence. A possible explanation came from the well plate assay: when three measurements were performed consecutively to check if the reaction product was prone to photo-bleaching, the fluorescence unexpectedly increased after each measurement. This suggested that the UV light used in the excitation scan promoted a photochemical step, which was required or beneficial for the final fluorophore. Therefore, the droplets were exposed to light of 365 nm for 10 min using a hand-held UV-lamp (*ca.* 100 µWcm⁻²). Indeed, the droplets containing HG3.17 were now red fluorescent, whereas those without enzyme were not (Figure 5.4A). This was the first indication, that there may be a two-step mechanism involved. Presumably, the enzyme first converted the substrate to the Kemp product, which then reacted further upon irradiation with UV light. A 1:1 mixture of the droplets after exposure to UV light did not show exchange of the red fluorophore after overnight incubation (Figure 5.4B).

In summary, the emergence of red fluorescence was observed both in droplets and well plates. Its appearance was accelerated by HG3.17 and was dependent on exposure to UV-light. The resulting fluorophore did not exchange between droplets. These observations are remarkable for four reasons:

- Only one fluorogenic substrate for the Kemp elimination was reported to date [229]. This substrate is based on the fluorophore coumarin, which is a bicyclic aromatic compound, is known to leak from droplets, and has excitation and emission bands in the UV range (300 to 450 nm) [230, 231].
- Assuming the structures of the respective products of **6a** and **7a** are the expected cyanophenols, the fluorescence would be far red-shifted for a small conjugated sys-

tem. Usually, larger π -systems are needed for excitation and emission bands in the visible light range [230].

- Another remarkable property is the large Stokes-shift. The Stokes-shift is the difference in the wavelengths of the excitation and emission maxima. In this case it is 100 nm. Usually, it is much smaller at about 20 nm [230], such as in the case of fluorescein with maximum absorption at 492 nm and maximum emission at 512 nm [109].
- Finally, the very low background activity in the absence of catalyst and the high sensitivity offered by a fluorescence readout compared to absorbance render these substrates attractive for studies of the Kemp elimination in general. In particular the smaller substrate **6a** is very similar to **2a** and the series of other 5-substituted Kemp substrates used to study linear free energy relationships.

Taking together these considerations, it was clear that these substrates were promising candidates to screen for Kemp eliminases. However, the observed fluorescence properties indicated that it was not the simple Kemp reaction product causing the fluorescence. The working hypothesis emerged, that there were two reactions involved. First, a dark reaction product, presumably the Kemp product, was generated. Second, a fluorescent reaction product or complex was obtained by exposure to UV light. Proving that the existence of these two reaction steps was the goal of the next section.

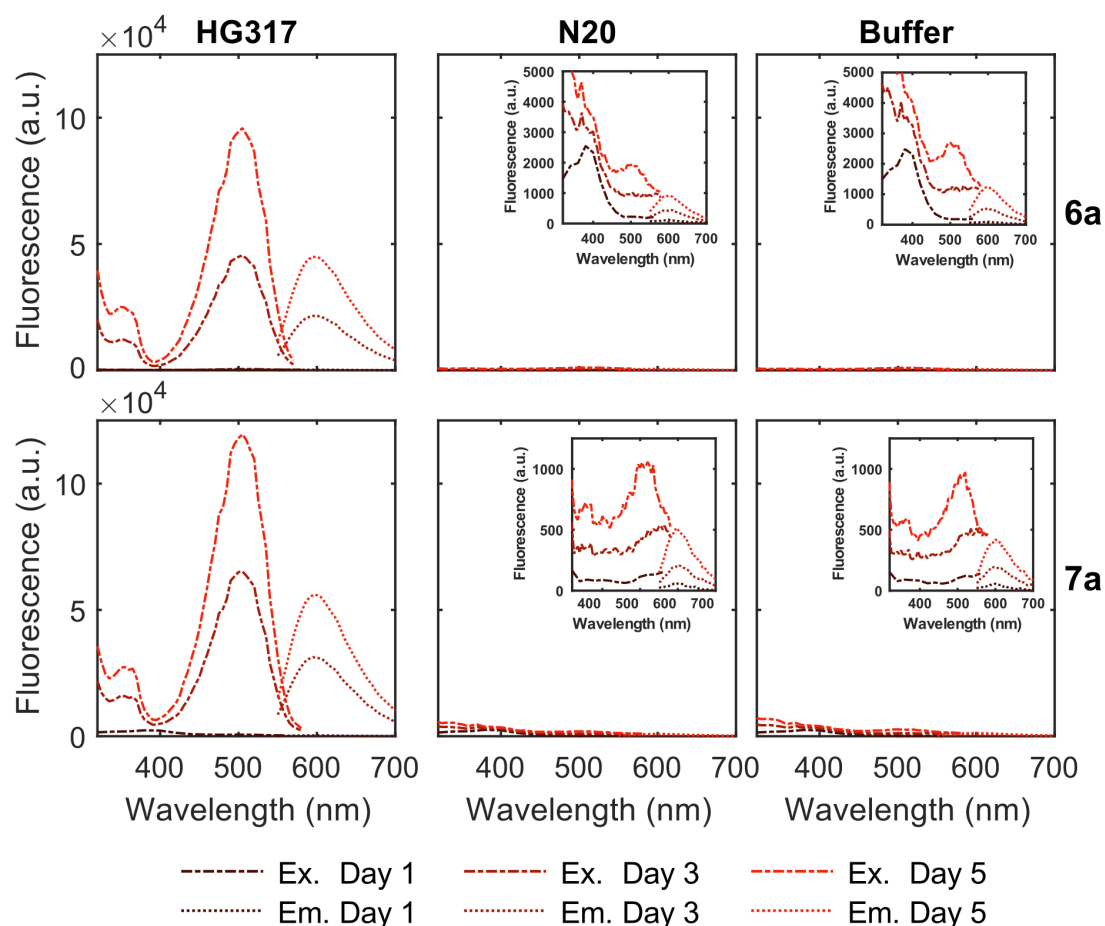


Fig. 5.3 Excitation and emission spectra for the reactions of substrate **6a** (top) and **7a** (bottom) with purified enzymes HG3.17, N20, and just buffer (20 mM Tris pH 7 and 50 mM NaCl), post UV-illumination. The red fluorescent product appeared spontaneously in buffer and in presence of N20, but its emergence was clearly promoted by Kemp eliminase HG3.17. There was an excitation maximum at 500 nm and an emission maximum at 600 nm. Excitation scan: emission measured at 620/20 nm; emission scan: excitation at 505/20 nm.

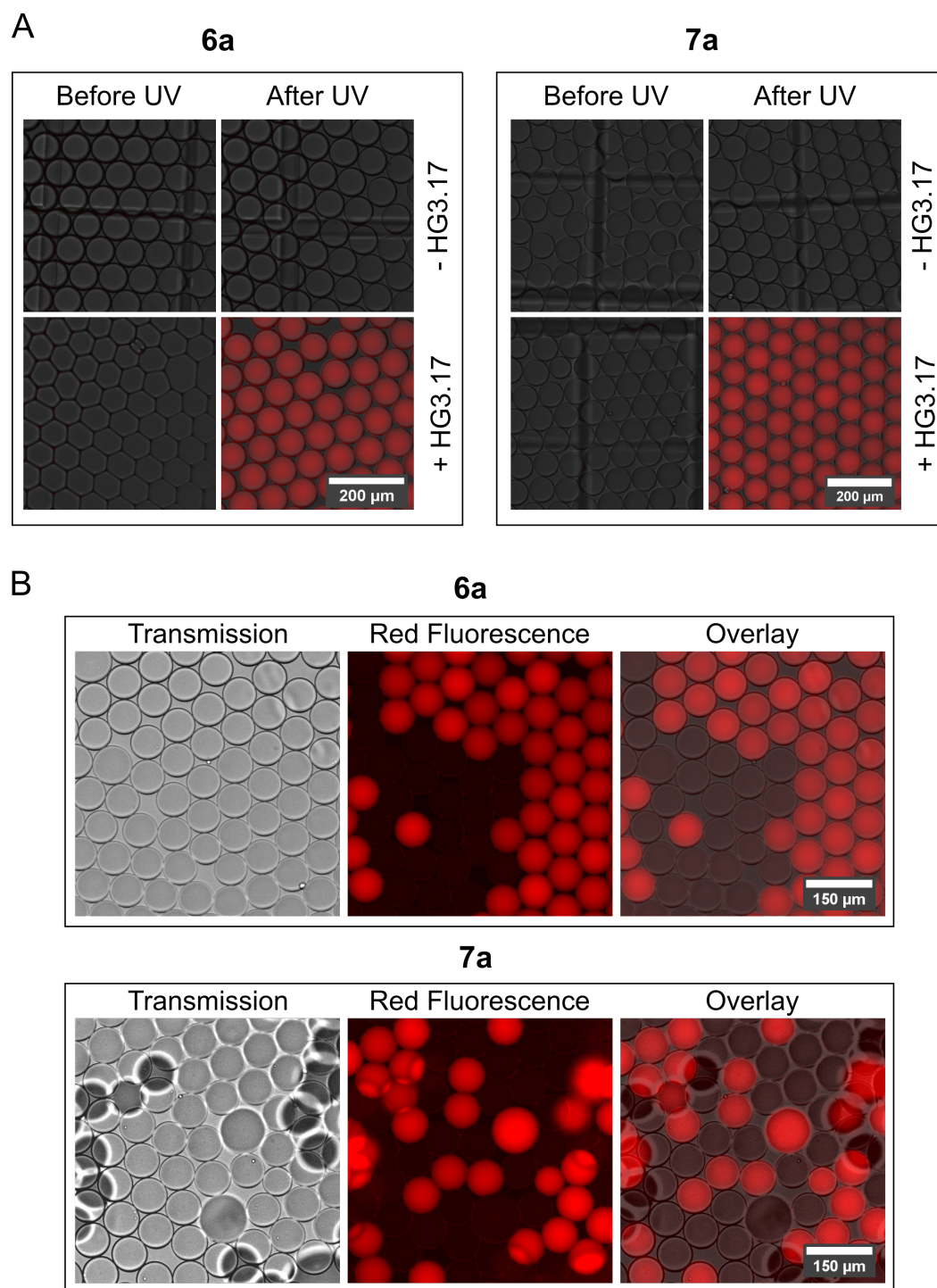


Fig. 5.4 A: After four days of incubation in darkness, red fluorescence did not appear for either substrate with or without enzyme. Exposure to 365 nm UV light for 10 min generated the red fluorophore in droplets that contained the enzyme HG3.17, but not in droplets with buffer only. Shown are the overlays of the transmission and red fluorescence images. B: A 1:1 mixture of the droplets with and without enzyme after exposure to UV light showed no exchange of the red fluorophore after overnight incubation.

5.2.2 Two reaction steps lead to fluorescence

By using absorbance instead of fluorescence scans, the two reaction steps could be distinguished. The reaction was performed in triplicate in a quartz plate for transparency in the UV range. Due to a simpler absorbance spectrum, only substrate **6a** was used for this characterisation.

The reaction was incubated for 24 h in the dark and monitored using absorbance scans from 280 to 600 nm. The absorbance increased at 350 nm while it decreased at 310 nm. Isosbestic points were observed at 290 and 330 nm (see Figure 5.6). The presence of isosbestic points indicates the conversion of one substrate to one product. Furthermore, the reaction followed pseudo first order kinetics as shown by plotting the change of absorbance against time at a single wavelength. The half life of the reaction was (4.3 ± 0.1) h as estimated by single-exponential fits at 310 and 350 nm, *i.e.* after 24 h of incubation the reaction was 98% complete.

After 24 h, the reaction was exposed to UV light of 365 nm for intervals of 5 s. There was an increase in absorbance at 300 and 500 nm. The appearance of the 500 nm absorbance band explains why fluorescence excitation at this wavelength is only possible after exposure to UV light. The absorbance at the other wavelengths did not change, apart from a decrease at 365 nm possibly due to photo-bleaching. Above a total exposure time of 30 s, the absorbance started decreasing at all wavelengths, indicating light-induced degradation of the compound. No isosbestic points were observed, but the absorbance spectra did not cross each other. Therefore it is not evident whether the intermediate is fully or only partially converted by the process.

The absorbance and fluorescence emission of product **6c** was dependent on pH. This could be due to protonation and deprotonation of the phenolic oxygen, which is expected in the Kemp reaction product. Upon addition of HCl the absorbance bands at 300, 350, and 500 nm diminished (Figure 5.7). A separately-prepared sample was diluted 1:9 with buffers

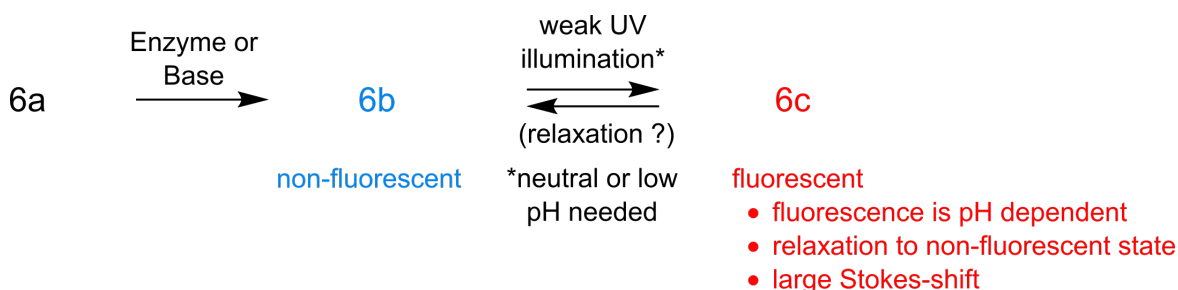


Fig. 5.5 Substrate **6a** was converted using enzyme or base to a non-fluorescent intermediate **6b**, which was further converted to fluorescent product **6c**.

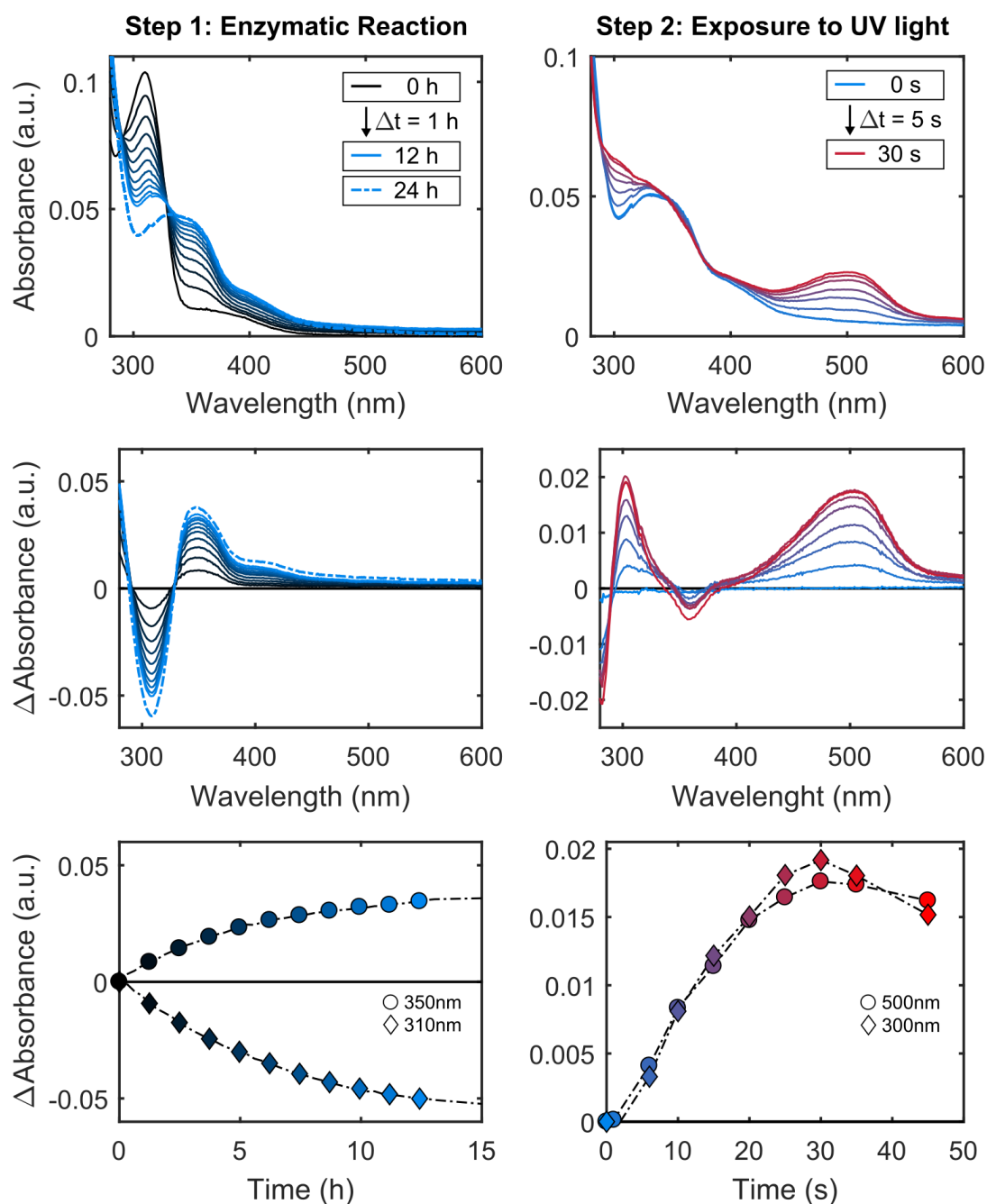


Fig. 5.6 The reaction of 1 mM **6a** with 1 μ M HG3.17 was followed by measuring absorbance spectra over time. The enzymatic reaction (Step 1) was performed in the dark and reached completion after 24 h. No increase of absorbance was observed at 500 nm during the enzymatic reaction. Only during exposure to light of 365 nm (Step 2), did an absorbance band at 500 nm appear, which can explain the gain of fluorescence upon excitation at this wavelength. The sample was irradiated with approximately $100 \mu\text{Wcm}^{-2}$; buffer: 20 mM Tris pH 7, 50 mM NaCl.

of varying pH. Fluorescence excitation spectra showed excitation bands at 300, 350, and 500 nm. The emission of fluorescence decreased below a pH of 6.5. The intensity was about half at pH 5, indicating the group which was protonated had a pK_a in this range. The effect was visible by eye under 365 nm UV-light (Figure 5.7D). Next, it was tested if the conversion of the intermediate to the fluorophore benefited from alkaline pH. The intermediate was added to either water or 1 mM NaOH and exposed to UV light. Interestingly, no fluorescence developed in the basic solution. When base was added to the neutral sample after UV exposure, the fluorescence increased. This indicated, that a proton was required for the conversion of the intermediate, but not for fluorescence in the final product.

Furthermore, the bands observed in the fluorescence excitation spectra matched the bands observed in the absorbance spectrum, indicating that they are part of the same compound, *i.e.* that exposure to UV light converts the majority of intermediate to final product.

Taken together these results established that a non-fluorescent intermediate was produced by the enzymatic reaction, followed by a non-enzymatic, UV-induced, and pH dependent conversion to the fluorophore. Next, the intermediate (**6b**) and the final product (**6c**) were separately prepared and purified for structural investigation.

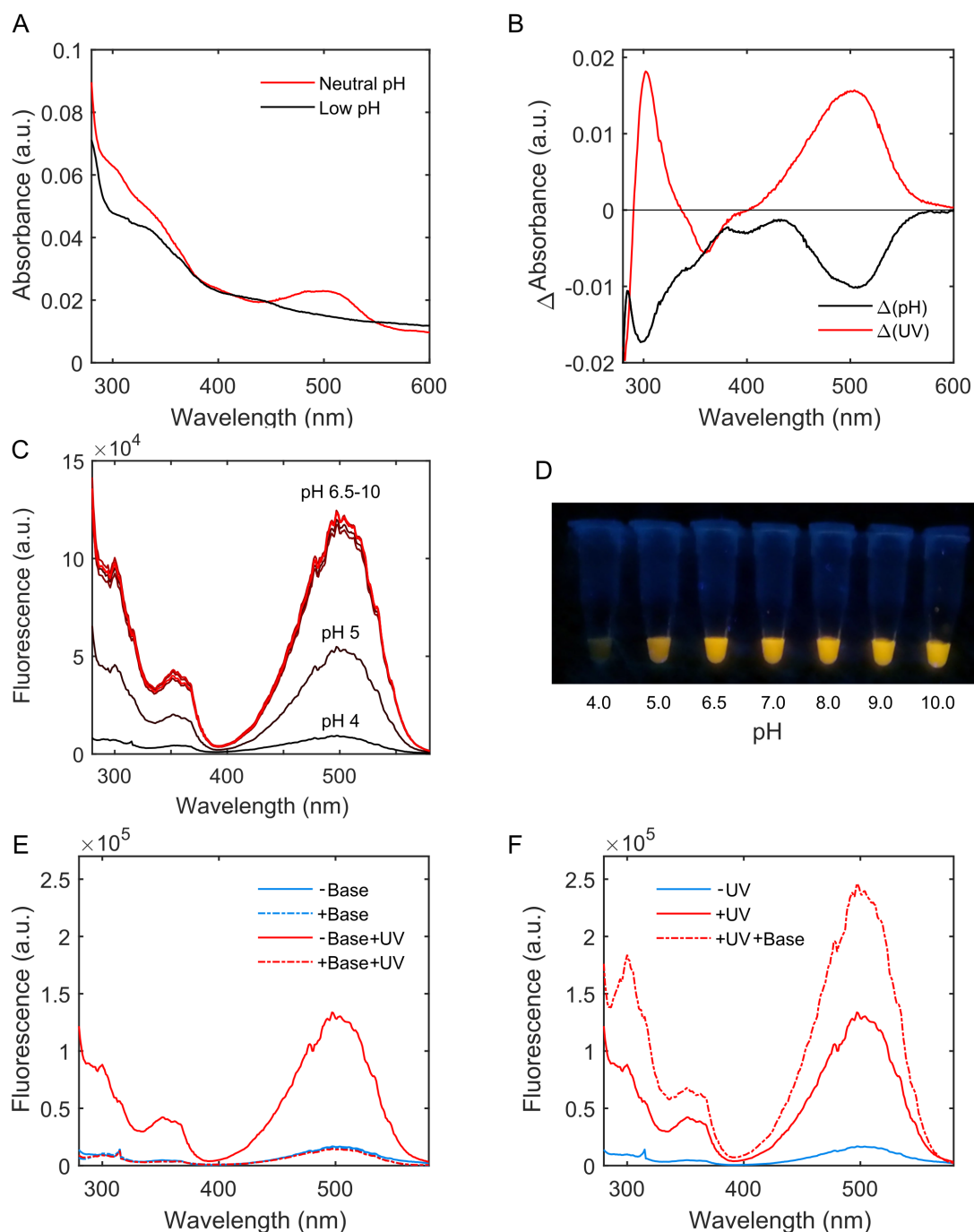


Fig. 5.7 The emission of red fluorescence by product **6c** was dependent on pH indicating the presence of at least one protonable group. *A&B*: Addition of HCl to the final reaction mixture quenched the absorbance bands gained in the previous experiment by exposure to UV light. *C&D*: **6c** was diluted into buffers of different pH. The fluorescence excitation spectra showed that below a pH of 6.5 emission was reduced. *E*: Basic conditions (1 mM NaOH in water) prior to illumination prevented the formation of the fluorophore, indicating that a proton is involved in its formation. *F*: Addition of base after conversion in unbuffered solution increased fluorescence emission. Buffers all at 20 mM with 50 mM NaCl; pH 4-5 sodium-acetate; pH 6.5-8.0 sodium-phosphate; pH 9-10 sodium-carbonate.

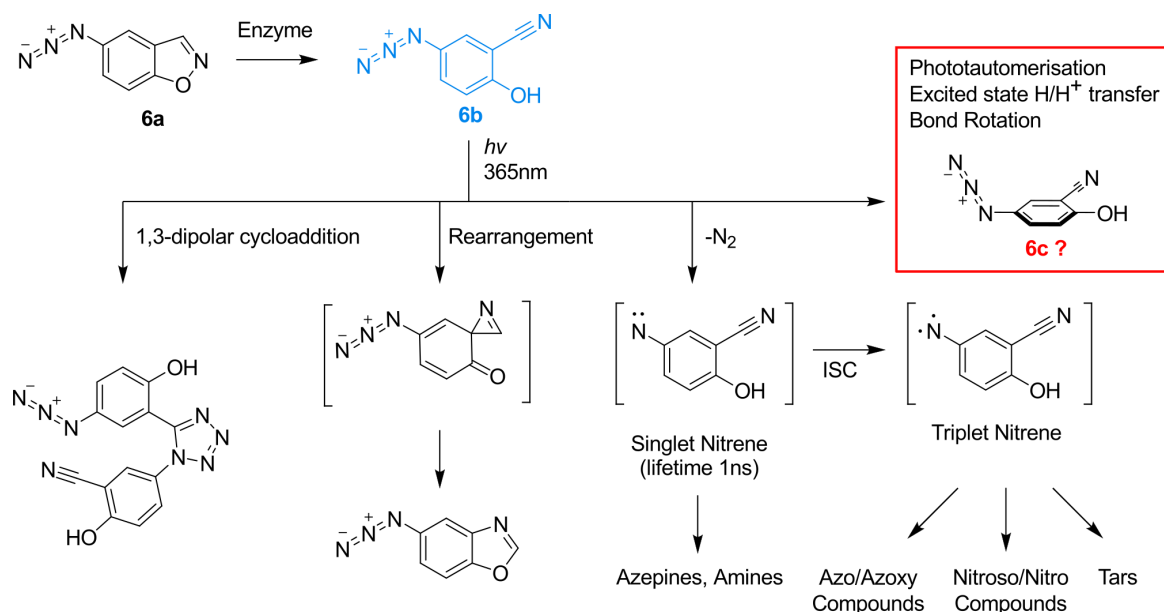


Fig. 5.8 Shown are possible reaction pathways for **6a**. ^1H -NMR indicated that the product of the first step is the expected Kemp reaction product **6b**. There are several possible reactions that could lead to the fluorophore (explained and in the main text), but the evidence pointed towards a small structural change (red box).

5.2.3 NMR and Mass spectroscopy of the reaction products

The NMR and mass spectrometric analysis confirmed the conversion of the substrate **6a** to the Kemp product **6b** in the first reaction step. The spectra of the main product in the second step, **6c**, were similar to those of **6b** as will be outlined.

First, the substrate was converted using enzyme HG3.17 and either kept in the dark or exposed to UV light. The enzyme was then precipitated using formic acid, the solutions lyophilised and the powders re-dissolved in dimethylsulfoxide (DMSO) or acetonitrile/water (1:1) for NMR and mass spectrometric analysis, respectively. The purification procedure did not degrade the red-fluorescent product, as dilution into aqueous solution and measurement showed the same fluorescence behaviour as before treatment. The reaction product of step 1 was not stable, *i.e.* developed fluorescence over time, even when kept in the dark. Thus, the analyses were performed directly after overnight incubation with the enzyme, which meant that the reaction had not always come to completion.

If **6b** was indeed the Kemp reaction product, the expected difference between the ^1H -NMR spectra was the loss of one of the four protons and an upfield shift (towards lower ppm) of the remaining three protons due to increased electron density at the oxygen. This was observed. The ^1H -NMR spectrum of the substrate **6a** showed the four expected protons and it was possible to unambiguously assign them using the coupling pattern and constants.

Referring to Figure 5.9, the proton furthest downfield at 9.20 ppm is the proton at position 3. The protons at positions 4, 6 and 7 are at 7.67, 7.41, and 7.83 ppm, respectively. The spectrum of **6b** showed that the reaction had been stopped at about 80% conversion of substrate to product. The intensities of the substrate peaks were reduced concomitantly and three new product peaks appeared. These were shifted upfield as expected. The proton at position 7 showed the strongest shift with -0.74 ppm consistent with the presence of the phenolic oxygen. The peak at 8.16 ppm is likely to stem from formate, which was used in the preparation of the sample. These observations are consistent with the expected Kemp reaction product.

With the ^1H -NMR spectrum of **6b** matching the expectations for the Kemp reaction product, it is of interest to consider what kind of reactions it could undergo. Four major possibilities emerged after a review of the literature shown in Figure 5.8.

First, azides undergo cycloaddition with cyanides, *i.e.* the reaction product could react with itself (or remaining starting material) [232–234], which would create a larger conjugated system. However, this type of reaction requires elevated temperature and is not known to be promoted by light. In the ^1H -NMR spectrum, the proton shifts between the two benzene moieties would be expected to differ, which was not observed. For these reasons, this reaction pathway was ruled out.

Second, 1,2-benzisoxazole photochemically rearrange to 2-hydroxynitrile [235, 236]. This can further rearrange, albeit at long illumination times, to benzoxazole [235]. The azide substituent may allow this process to proceed faster. However, this product would lack a protonable functional group and thus be inconsistent with the observed pH dependent emission of the fluorophore.

Third, aryl-azides are known to be photo-sensitive and are prominently used for bioconjugation [237]. Upon irradiation, N_2 is eliminated creating a singlet nitrene which rapidly converts to a triplet nitrene [238, 239]. The singlet nitrene can react to form amines, but typically self-inserts into the ring and reacts with a nucleophile to form azepines [237, 239]. The triplet nitrene can dimerise to form azo- or azoxy compounds, oxidise to nitroso or nitro compounds, or polymerise to form insoluble tars [238, 239]. Which reaction path is favoured depends highly on the ring substituents. Using this chemistry, Lord *et al.* synthesised a series of azide-based, photo-activatable fluorophores. The non-fluorescent azide was illuminated at 407 nm, which caused conversion to the amine. The amine exhibited red fluorescence, which was explained with an electron push-pull mechanism [240]. While this constitutes an interesting precedent, the cleavage of the azido group required $10\times$ stronger irradiation for $10\times$ as long in the cited example. Cleavage of the azido group would also be inconsistent with the fluorescence of the products of substrate **7a**. In this substrate, the azide was converted to a tetrazole before irradiation. In addition, none of the predicted masses or NMR shifts of

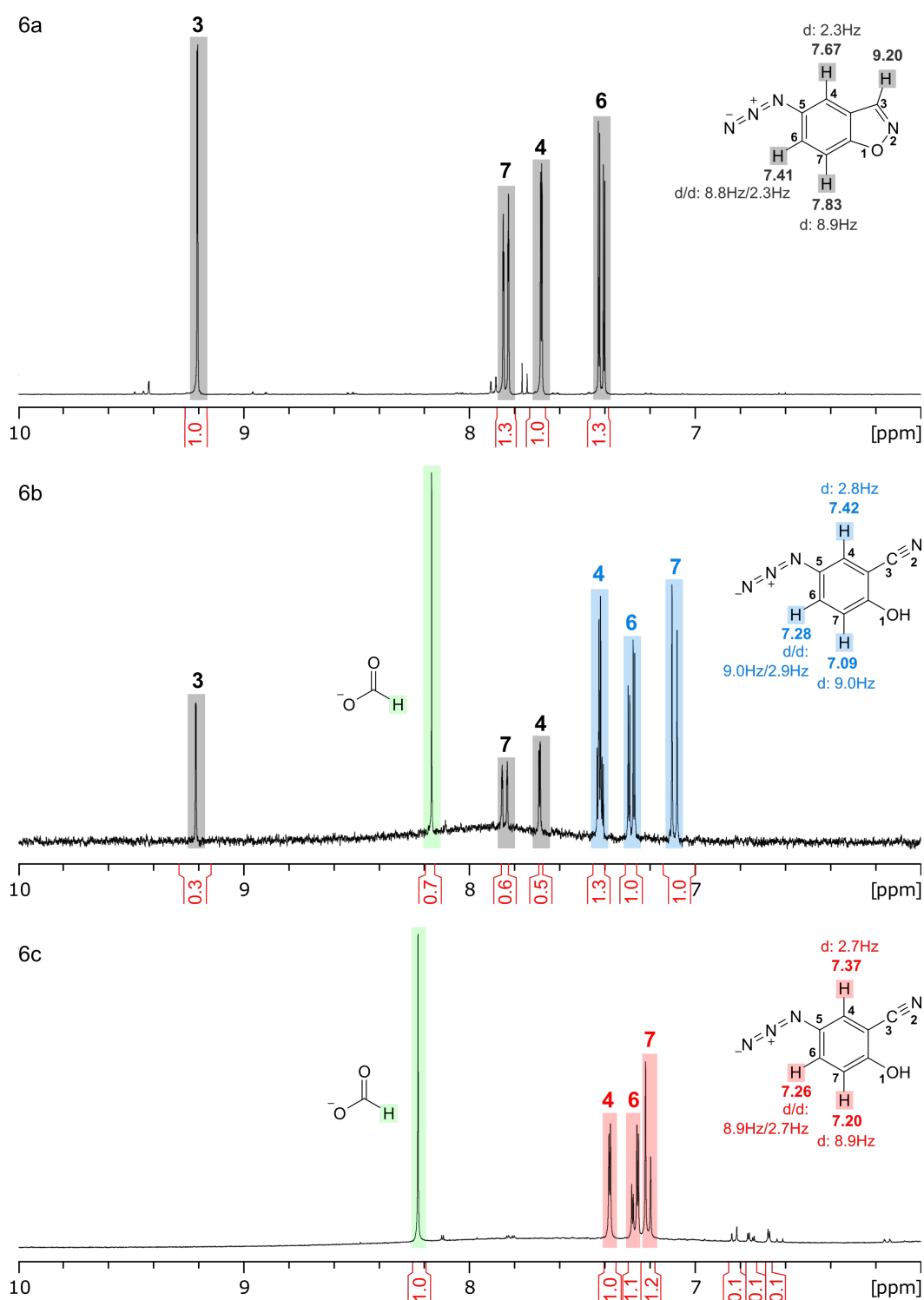


Fig. 5.9 ^1H -NMR spectra for **6a** (substrate, grey), the dark reaction product (**6b**, blue), and the product after UV exposure (**6c**, red). Formate (green) was a contaminant peak introduced during the purification procedure. As expected, one proton is lost during the dark reaction and the remaining three shift upfield. Interestingly, the ^1H -NMR of **6c** revealed only tighter grouping of the three proton peaks compared to **6b**, indicating no major structural change of the majority component.

the possible reaction products (*e.g.* the diazo-compound) were observed¹. Therefore, this pathway was excluded as well.

Finally, a subtler change in molecular structure may occur, *e.g.* due to phototautomerisation, irreversible excited state H or H⁺ transfer [241], or the rotation of the azide group from *trans* to *cis* with respect to the cyano group, or *vice versa*, or out of plane as shown in Figure 5.8. This could enable a previously inaccessible electron transition and explain the observed disappearance of fluorescence by thermal relaxation. The rotation could also enable the formation of excimers or exciplexes. These are excited state complexes of a molecule with itself or another compound, which show red-shifted fluorescence emission compared to the monomer and therefore a large Stokes-shift [230]. This and a reversible excited state H or H⁺ could explain fluorescence emission at high wavelengths by a small conjugated system [242].

The ¹H-NMR spectrum of **6c** did not show any new peaks. The three proton peaks observed for **6b** shifted closer together, the distance between proton 4 and 7 diminishing from 0.33 to 0.17 ppm. This small shift hinted at only a subtle change of the structure. Several samples were prepared and combined to obtain a higher concentration of product to observe possible side products, which could be responsible for the fluorescence. The integral of the peak which was previously assigned to formate now matched the integral of the three protons. This could be consistent with a benzoxazole reaction product (Figure 5.8, product of rearrangement). However, the ¹³C, DEPT, HSQC, and HSQC NMR spectra showed that the proton at 8.22 was indeed part of formate and confirmed that the principal compound in the sample remained consistent with the Kemp reaction product (Appendix Figure D.1 to D.4). The carbon atoms adjacent to the hydroxy, cyano and azido substituents had typical chemical shifts of 158, 100, and 131 ppm, respectively. Three minor peaks in the ¹H-NMR between 6.6 and 6.9 ppm showed the coupling pattern expected for a 1,2,4-trisubstituted benzene derivative. The integrals suggested that this product could constitute 8% of the final mixture. The chemical shifts could be consistent with the amine product, whose mass was however not found in the mass spectrum. If any self-complexation processes played a role in the gain of fluorescence in the final step, *e.g.* by π - π stacking, it was possible that this process would not occur in DMSO. Therefore, it would have been desirable to perform ¹H-NMR using a dilution series of the compound in D₂O. However, this was not possible due to the limited solubility of the final compound in water.

To investigate the presence of any other reaction products not detectable by NMR, analytical reversed-phase high-performance liquid chromatography (RP-HPLC) of the samples was performed, see Figure 5.10. Importantly, only one major peak was observed for **6c** and

¹The molecular mass and ¹H-NMR spectrum for all the products of the above pathways were predicted using the program ChemDraw and compared to the experimental data.

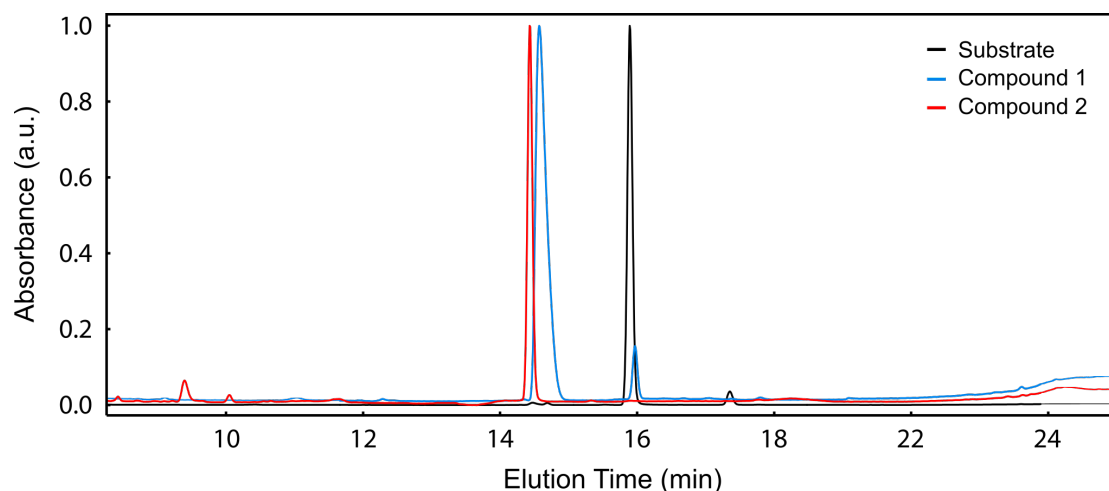


Fig. 5.10 RP-HPLC analysis of **6a** (substrate, black) and the dark (**6b**, blue) and UV reaction products (**6c**). **6b** eluted almost 2 min earlier than the substrate as expected for a decrease in hydrophobicity. As in the NMR experiment, only one major product is observed and there is only a small difference between **6b** and 2. The main peak was associated with red fluorescence, ruling out that any of the minor peaks could be responsible for it.

it was associated with red fluorescence after fraction collection of this peak. A small shift to earlier elution time was observed for **6c** compared to **6b**, suggesting a small increase in hydrophilicity. This shift is 10× smaller compared to the shift observed between the substrate and **6b**.

Liquid chromatography-mass spectrometry (LC-MS) of **6c** provides further evidence that the final product is similar to the Kemp reaction product. The main peak was consistent with the $[M-H]^-$ ion (m/z for $C_7H_3N_4O^-$: 159.03 $[M-H]^-$; found: 159.2), see Figure 5.11. All ions detected in this peak were consistent with fragments of the Kemp reaction product. A smaller peak eluted earlier and showed an m/z of 163.2 which is consistent with the $[M-H]^-$ ion of 5-nitro-2-hydroxy-benzonitrile (m/z for $C_7H_3N_2O_3^-$: 163.01 $[M-H]^-$; found: 163.2). The minor peaks observed in the 1H -NMR spectrum of **6c** were not consistent with 5-nitro-2-hydroxy-benzonitrile, as these were previously determined to be at 8.57, 8.33, and 7.13 ppm respectively. This suggests that 5-nitro-2-hydroxy-benzonitrile is a minor reaction product. However, it does show that elimination of N_2 from the azide and oxidation to the nitro group was a possible reaction path. The nitro-compound by itself is not known to be fluorescent and mixtures with either the substrate or **6b** did not yield the fluorescence before or after illumination.

Given the detection of at least one product which lost the azido functionality, its integrity in the bulk component of the final reaction mixture was tested by reacting it with the same DBCO-derivative used to synthesise substrate **7a**. Indeed, after the reaction only the expected

mass was found and the compound at m/z of 159.2 was not observed (m/z for $C_{36}H_{39}N_6O_7^-$: 667.29 $[M-H]^-$; found: 667.7).

Summarising all of the above, the observations using NMR and mass spectroscopy showed that the structure and mass of the main reaction product both pre- and post-illumination was consistent with the Kemp elimination product **6b**. The difference pre- and post-illumination of the main component in the mixture is small. Yet, RP-HPLC analysis of the post-illumination sample linked the observed fluorescence with the main component in the mixture. Together these data suggest, that gain in fluorescence is achieved not by a major structural change such as a dimerisation, but by a more subtle process such as the rotation around the nitrogen-carbon bond of the azido compound as suggested in Figure 5.8. This rotation could have an energy barrier due to partial double-bond character of the nitrogen-carbon bond and be overcome by the absorption of a photon. Further investigation is required to determine if this is a likely origin of fluorescence. For example, *ab initio* calculation of the energies of the highest occupied and the lowest unoccupied molecular orbital for different geometric orientations of the azido group would indicate if there is a decrease in the energy gap, which would explain the observed fluorescence. The main reaction product after both reaction steps was consistent with the elimination of a proton from the 3 position in **6a**, *i.e.* the Kemp elimination. Therefore, this substrate was further investigated for use in droplet screening.

5.2.4 The single-cell lysate assay yields green fluorescence

The reaction in droplets was shown above using purified enzyme. Next, it was tested if the same results could be obtained using single-cell lysates in droplets. Also, since the readout was now fluorescence, smaller droplets of 15 μm diameter were used. These droplets have the advantage that the cell lysate is diluted 100 \times less compared to 70 μm droplets. Combined with the higher sensitivity of fluorescence measurements compared to absorbance, this should offer a vastly more sensitive assay than used in the previous section.

Fluorescence was not observed in droplets containing single-cells expressing HG3.17 using the original buffer conditions and UV exposure after overnight incubation. Bulk measurements showed that the reaction stalled after 30 min in Tris buffer, whereas fluorescence kept increasing in NaPi buffer over the duration of the experiment (12 h). Changing the buffer in the droplet experiment still did not yield fluorescence after overnight incubation followed by UV exposure. Taking into account that no leakage was observed for **6c**, it was speculated that **6b** leaked out of droplets (this could for example be explained by a reduction of the pK_a).

Therefore, an experiment was performed in which the droplets were generated and immediately exposed to 365 nm light for 30 s every 15 min. Droplets were mixed before every

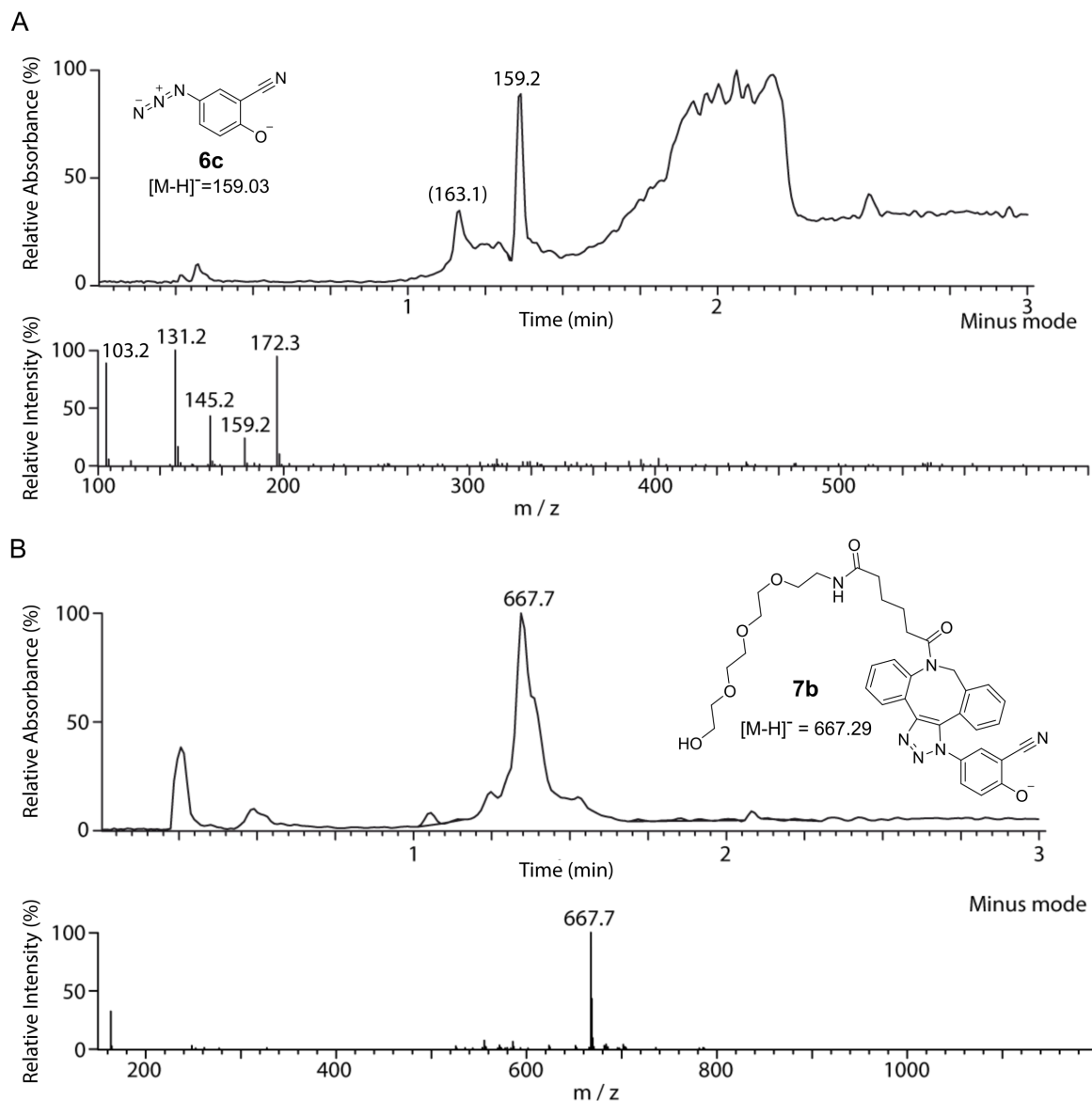


Fig. 5.11 A: The mass of **6c** was confirmed by LC-MS (m/z for $C_7H_3N_4O^-$: 159.03 $[M-H]^-$; found: 159.2). The other observed ions can be explained by fragmentation: 159.1 $[M-H]^-$, 145.2 $[M-H-N]^-$, 131.2 $[M-H-N_2]^-$, 103.2 $[M-H-N_2-CO]^-$, and 172.3 as an adduct of $[M-H-N_2]^-$ and acetonitrile (the solvent). B: The integrity of the azide group in **6c** was shown *via* a click reaction to yield **7b**. The compound was reacted with a DBCO derivative to yield the indicated structure and its generation was confirmed (m/z for $C_{36}H_{39}N_6O_7^-$: 667.29 $[M-H]^-$; found: 667.7).

exposure by slowly rotating the droplet container using a servo motor. The set-up was run using an Arduino micro-controller. This procedure finally yielded fluorescent droplets after overnight incubation. Interestingly, under these conditions fluorescence was only observed in the green channel, but not in the red channel of the fluorescence microscope (excitation: 470/22 nm and 531/40 nm, emission: 525/50 nm. and 593/40 nm). The experiment was repeated three times, with the same result. The intensity and wavelength at which fluorophores emit light is dependent on many factors such as solvent, temperature, viscosity, or specific interactions with other molecules [243]. Therefore, the influence of the reaction components and different solvents was investigated.

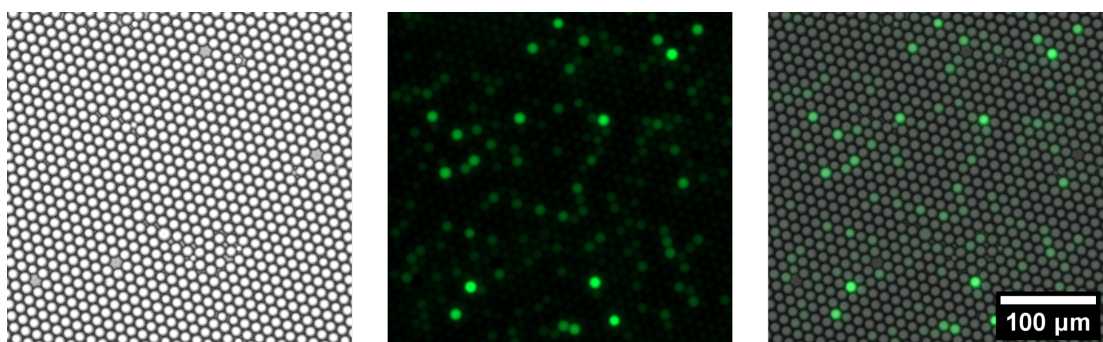


Fig. 5.12 When single cells expressing HG3.17 were encapsulated ($\lambda=0.2$) with **6a** and lysis agent in droplets, the reaction yielded green fluorescence after overnight incubation with periodic UV illumination. Conditions at reaction start: 1 mM **6a**, 20 mM NaPi pH 7.0, 50 mM NaCl, 0.1x BugBuster (lysis agent), 54 to 66 U/ μ L rLysozyme. Excitation 470/22 nm, emission: 525/50 nm.

5.2.5 The fluorophore emits at 335, 544, and 600 nm

The influence of the reaction components in the single-cell lysate assay and different solvents was investigated revealed that there are three possible fluorescence emission bands at 335, 544, and 595 to 605 nm respectively.

First, the effect of the lysis agent and the cell lysate was tested. A dilution series of lysis agent showed there was no change in the maximum emission wavelength, see Figure 5.13. However, an old sample of **6c** was used (about 1 week old, stored at room temperature in DMSO), which appeared to have lost its fluorescence (from 10^4 down to 10^3 fluorescence units using the same instrument settings). After exposing the dilution series to 365 nm light, the fluorescence was recovered. This indicated that the formation of the fluorophore was reversible and had possibly undergone thermal relaxation during storage. Furthermore, this experiment showed that the lysis agent did not interfere with the formation of the fluorophore.

In contrast to the lysis agent alone, the cell lysate of *E. coli* cells altered the emission of the fluorophore. At 100% cell lysate, which corresponds to a 50 to 100 fold dilution of the cytosol, the emission at 600 nm was 12 fold reduced. At the same time, a new peak at 335 nm appeared. In the absence of cell lysate, there was a minor peak at 310 nm, which was most likely the Raman peak due to scattering by water [244]. In 15 μm droplets, the cytosol of a single cell is diluted about 2,000 fold. Therefore, the emission at 335 nm should not play a major role.

Because this did not explain emission of green light, the fluorescent dye was dissolved in a series of different solvents to see if the fluorescence could be shifted to green emission. Indeed, in all solvents initially tested the emission maximum shifted from 600 nm to 540-545 nm, see Table 5.1. The amount of fluorescence emitted, as measured by area under the emission curve, was 10 to 100 \times less in any solvent compared to water. In trying to understand the shift in emission maximum, the properties of the different solvents were compared. Kamlet *et al.* developed one empirical set of parameters derived from the solvent effects on many fluorophores. It takes into account solvent dipolarity (π^*) as well as hydrogen-bond acceptor (α) and donor (β) strength [245]. Comparing the different values, it appeared that a high π^* but low β value favoured emission at a higher wavelength. The first is expected. When a fluorophore goes to the excited state, its dipole typically increases. In response, the solvent molecules surrounding it reorganise, which lowers the energy of the excited state causing a red-shift of the emission. The more polarised the solvent, the stronger the stabilising effect and therefore the red-shift [230, 243]. The second observation is consistent with the observation that protonation of the fluorescent dye quenches fluorescence.

If the shift to green emission was due to solvent relaxation alone, then a smooth shift from 600 to 540 nm would have been expected in mixtures of water and other solvents. Instead, the existence of two emission bands was observed, Figure 5.14. This is indicative of emission from two different excited states, *e.g.* a locally excited state and an internal charge transfer state which are differentially stabilised by the different solvents [243]. The interaction with nitrobenzene appeared to be specific rather than based on general solvent properties, because even a small percentage of solvent locked the emission into a broad band.

In summary, it was found that **6c** exhibits complex fluorescence behaviour with at least three emission bands at 335, 544, and 600 nm depending on solvent conditions. These insights indicate, that the green fluorescence observed for the single-cell lysate assay in droplets originates from the same compound which was characterised previously.

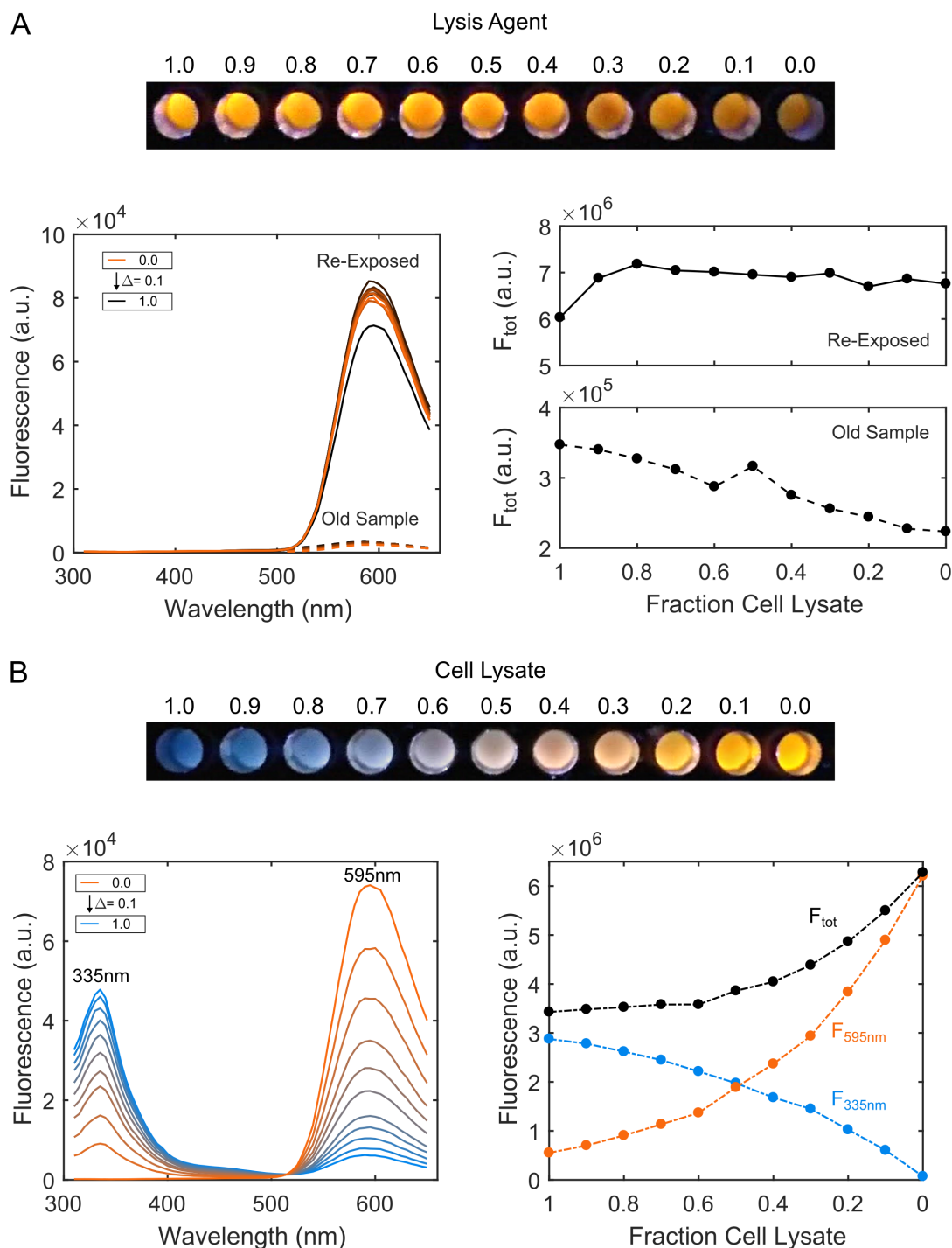


Fig. 5.13 The effect of additives on light emission by **6c**. **A**: Dilution series of **6c** with lysis agent (BugBuster, Merck-Millipore) starting from the recommended concentration (1.0). An old sample of **6c** was used and upon *in situ* re-exposure to UV light of 365 nm, strong fluorescence was recovered. **B**: Dilution series of cell lysate (1.0 corresponds to 50-100 fold dilution of the cytosol and contains 1.0x lysis agent). The cell lysate quenched fluorescence at 595 nm while a new peak at 335 nm appeared. Conditions: 1 mM **6c** in 20 mM NaPi pH 7.0, 50 mM NaCl, images of wells illuminated at 365 nm.

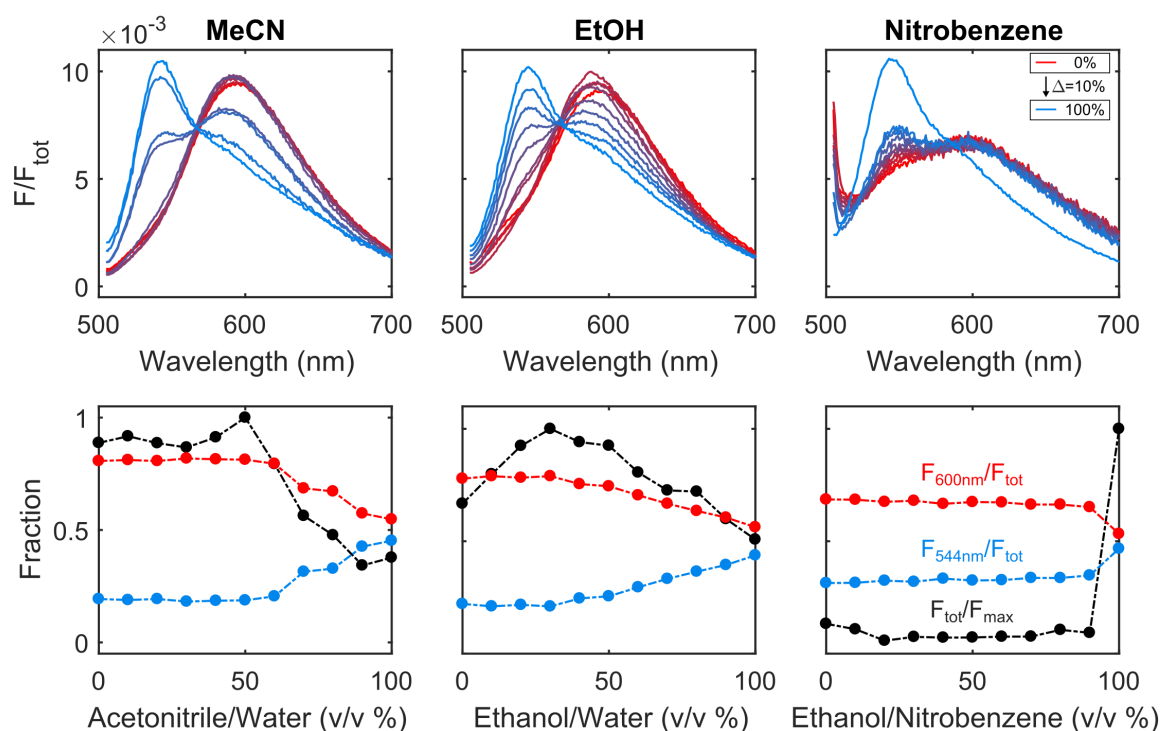


Fig. 5.14 Solvents affect the fluorescence emission of **6c**. The top row shows the emission spectra for the different solvent mixtures; the bottom row shows the fraction of the total emission if acetonitrile (MeCN) or ethanol (EtOH) were added to water an emission band at 544 nm emerged while the emission at 600 nm decreased. Total emission of fluorescence between 505 and 700 nm was reduced by half in acetonitrile and methanol compared to water. Fluorescence emission was broad in nitrobenzene and 20 \times reduced compared to water. Even at 90% v/v ethanol emission was strongly reduced, but the band at 544 nm emerged. F_{tot} : area under the curve, F_{max} : highest F_{tot} , $F_{544\text{nm}}$ and $F_{600\text{nm}}$: area under the curve up to and above 567 nm (isosbestic point). Excitation at 470 nm.

Table 5.1 The maximum fluorescence emission wavelength in different solvents and the solvent properties.

Solvent	λ_{\max} (nm)	Solvent Properties [†]		
		π^*	α	β
Water	590	1.09	1.17	0.18
Nitrobenzene	596	1.01	0.39	0.00
Dimethylsulfoxide	540	1.00	0.00	0.76
<i>N,N</i> -Dimethylformamide	545	0.88	0.00	0.69
Acetonitrile	540	0.75	0.19	0.39
Tetrahydrofuran	543	0.58	0.00	0.55
Ethylacetate	540	0.55	0.00	0.45
Ethanol	540	0.54	0.83	0.77
Toluene	insoluble	0.00	0.00	0.00
Cyclohexane	insoluble	0.00	0.00	0.00

[†] values from reference [245].

5.2.6 Enrichment of HG3.17 using substrate 6a

Because the appearance of green fluorescence was reproducible and the number of green fluorescent droplets matched the expectation for a λ of 0.2, an enrichment experiment with enzymes HG3.17 and N20 was performed. The same phenotypic assay as for the AADS assay was used (Section 4.4.6). Droplets were generated with a λ of 0.2 and a starting ratio of 1:500 of positive to negative cells ($\epsilon_0=0.002$). The histogram of fluorescence after overnight incubation and intermittent UV exposure is shown in Figure 5.15. Negative droplets did not show a detectable fluorescence signal. The number of sorted droplets was therefore estimated by regularly determining the droplet frequency *via* high-speed video analysis and averaging over time. Thus, an estimated 10^6 droplets were sorted, out of which all droplets with a signal 2 fold above the detection threshold were collected. This yielded 392 droplets, close to the expected 400. Re-transformation of a third of the recovered DNA yielded 229 without a halo (HG3.17) and 26 colonies with a halo (N20, false positives). This translates to a recovery rate of two colonies per droplet and a post-sorting ratio of $\epsilon_1=0.9$, which is an enrichment of $\epsilon_1/\epsilon_0 = 450$. In summary, this experiment showed that the observed green fluorescence is functionally linked to the activity of HG3.17. This implied that the assay could be tested with the metagenomic library SCV for functional screening.

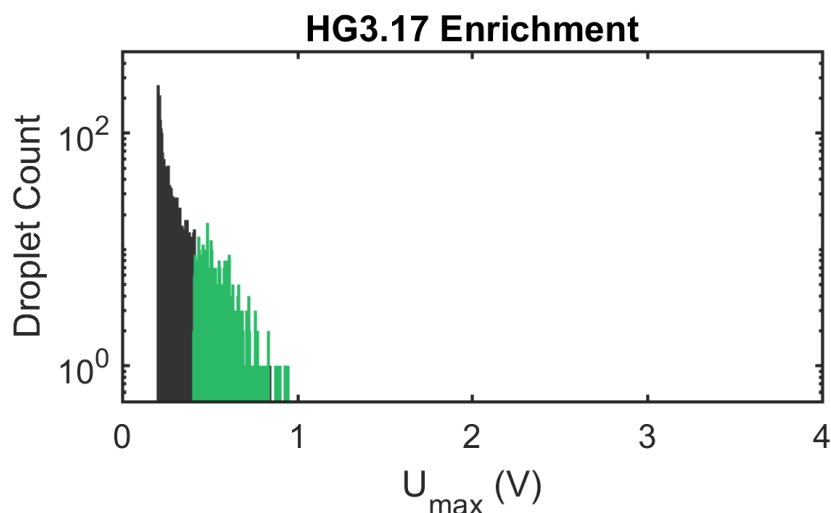


Fig. 5.15 Histogram of the enrichment of HG3.17 diluted 1:500 with N20 using the **6a** substrate. The droplets were incubated overnight and intermittently exposed to UV light of 365 nm. The majority of droplets did not give a signal. Their number was estimated at 10^6 using video analysis. Droplets with a signal 2 fold above detection threshold were collected (green bars). No droplets with a signal higher than 1 V were observed.

5.3 Functional metagenomic screening for Kemp eliminases

5.3.1 Screening of the metagenomic SCV library

The SCV library was prepared as for previous sorting campaigns and encapsulated at a λ of 0.2. The λ was reduced from the 0.35 used in the esterase screening campaign to reduce false positives due to co-encapsulation and erroneous sorting [60]. A total of 8×10^6 droplets were sorted over three days. The droplets were intermittently exposed to UV light over night, but not during the sorting periods (*ca.* 5 h). Histograms were obtained that resembled those obtained during the successful esterase screening campaigns (compare to Figure 3.11). That is, there were rare events of high fluorescence above the sorting threshold indicating the detection of Kemp eliminase activity. The average apparent hit rate was 2.1×10^{-4} compared to 4.3×10^{-5} and 4.7×10^{-5} during the esterase sorts. A direct, quantitative comparison of the hit rates between the two reactions is limited because of differing chemical and biological background rates. However, qualitatively they are both in the expected range of 10^{-5} to 10^{-4} for metagenomic screening of common enzymes. This finding is remarkable, because the data suggests that Kemp eliminases are occurring as frequently as esterases. The Kemp elimination has in general been seen as a *non-natural* reaction. To date, the highest promiscuous activity of an enzyme towards **2a** is in the range of $10^4 \text{ M}^{-1} \text{ s}^{-1}$ [201]. In the histogram of sorting day 1, there are 29 events with a signal higher than 1 V, which is above the signal

any droplet containing well-expressed HG3.17 achieved – the best Kemp eliminase known to date. Taking into account the low expression levels in metagenomics, this indicates that the observed events were caused by very active Kemp eliminases.

Table 5.2 Summary of the SCV library sort using **6a**.

Day	Droplets		Hit Rate [†] /10 ⁻⁴
	Sorted /10 ⁶	Collected	
1	3.3	211	3.2
2	3.9	105	1.4
3	0.8	22	1.2
Total	8.0	338	2.1

[†] expressed as hits per library member, assuming activity of each collected droplet was due to a single library member and λ of 0.2 for the sorted droplets.

To link the observed activities to a gene, the DNA of the collected droplets needed to be recovered, re-transformed, and re-screened. A total of 338 droplets was collected from which 5×10^4 colonies were recovered. This number was clearly above the expectation assuming typically observed recovery efficiencies of 2 to 5 colonies per droplet. Following the calculations of Baret *et al.* [60], one can estimate the carry-over of negative cells due to co-encapsulation. Using the apparent hit rate and occupancy $\lambda = 0.2$, one would expect 67 co-encapsulated cells. In the absence of biological bias, this would introduce only 17% false positives. In addition to co-encapsulation, random sorting errors due to flow-fluctuations dust particles can contribute to false positives, this error-rate is typically 10^{-4} or less. To explain all of the observed false positives, an apparent error-rate of 10^{-2} would be required; a rate which would have been observed during the droplet sort ². Therefore, it must be concluded that there was a bias towards false positives. Its origin is analysed in more detail in the next section.

A similar result had been obtained for the esterase screening campaign, albeit with a lower apparent error-rate of 10^{-3} . A secondary screen on culture plate was successful at identifying true positives. With the same goal in mind, 1,000 colonies were picked and grown in 12 96-well plates, 5 plates each for day 1 and 2 and two plates for the last day of sorting. After lysis, the lysates were screened at 2 fold dilution in 384-well plates using the substrate **6a**. However,

²The error-rate during a sort can be estimated by high-speed video analysis. Typically, a few thousand droplets are analysed in one recording, thus an error rate above 10^{-3} would have been noticed during the experiment. An error rate of 10^{-4} or lower may have gone unnoticed.

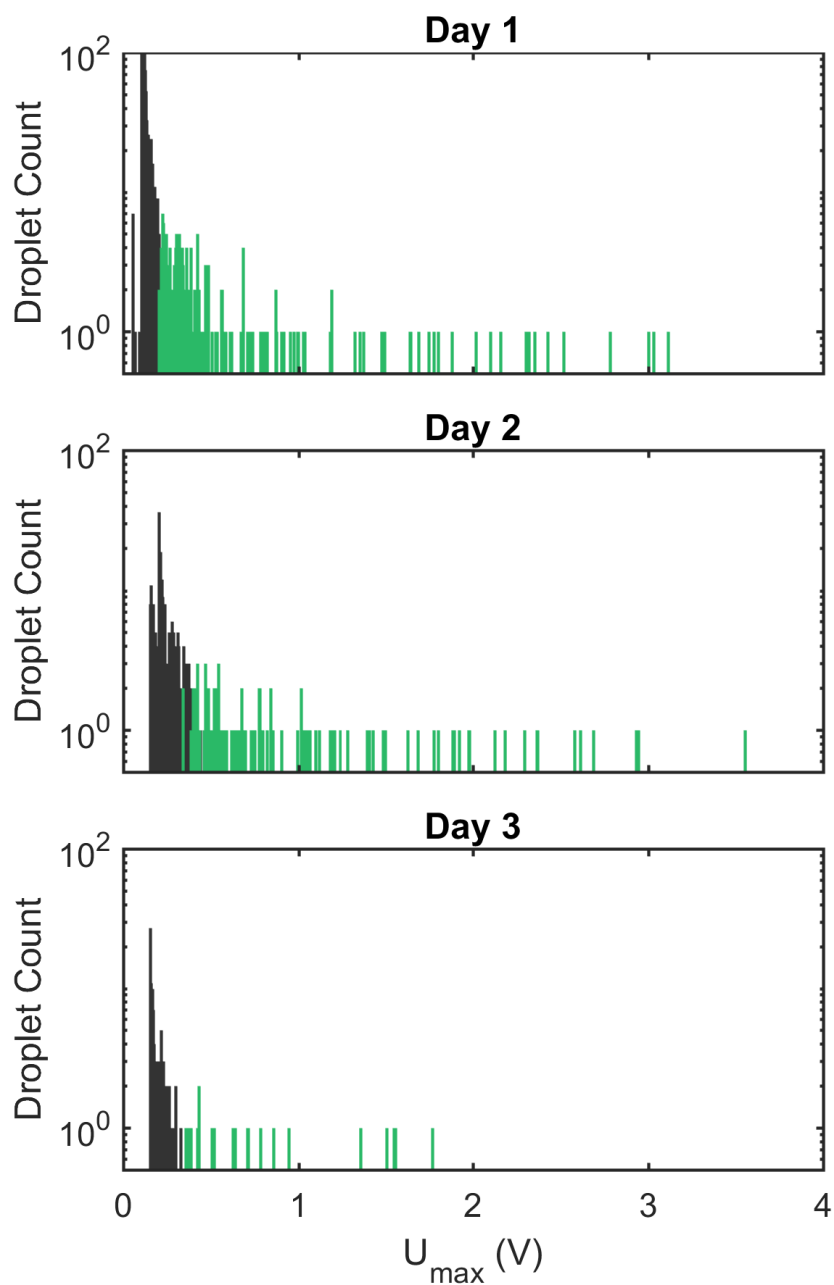


Fig. 5.16 Shown are the histograms of three droplet sorts of the metagenomic library SCV. The library was co-encapsulated with substrate **6a**, incubated, and sorted over three days with intermittent UV exposure during the overnight incubation periods. Excitation: 488nm, emission: 525/28nm.

out of the tested 1,000 colonies the signal of only 10 was just beyond three standard deviations above the average signal, *i.e.* there were no stand-out positives above background as shown in Figure 5.17. An emission spectrum recorded one week after incubation (sealed, in the dark) showed mixed ratios of emission at 335 nm and 595 nm indicating different degrees of quenching by the cell lysate. The variation in cell lysate concentration was likely to mask differences in product concentration. Nonetheless, to assess the sorting outcome, 30 variants were selected based on having 1) a positive change in fluorescence at each measured time-point and 2) having the highest overall change in fluorescence.

A repeat experiment at 10 fold dilution of the cell lysate, which should have shifted the majority of the emission to 595 nm, did not improve the outcome. Neither yielded a plate screen using the substrate **2a** any hits above background.

This showcases the superior performance of the droplet assay in detecting weak activities. In the above case, it allowed a 1,800 fold dilution of the cytosolic content of a single cell, while the accumulation of fluorescent product was confined to a volume of 1.8 pL. In this volume, just 1,000 product molecules would yield a concentration of 1 nM. To reach the same product concentration in the 384-well plate assay (10 μ L), 6×10^9 molecules would be required. For this reason, it was unsurprising that the well-plate assay did not yield clear hits in the way the droplet assay did. Furthermore, it was practically more difficult to ensure even illumination of the entire 384-well plate. For comparison, all of the sorted droplets together took up a volume of less than 10 μ L, *i.e.* would have fitted into a single well of the plate. In the absence of an immediate alternative, it was therefore important to identify the origin of the false positives in order to minimise their number in a new droplet screen. Ideally to such an extent that most recovered colonies could be assumed to be genuine hits.

5.3.2 Sequence analysis and origin of the false positives

The reason for the high number of false positives was found by sequence analysis of the selected hits. All but one of the 30 selected variants contained a compromised vector. The insert of the variant with intact vector was 472 bp in length. One complete ORF with a length of 300 bp was predicted, but a database search of its amino-acid sequence showed no homology with any previously sequenced gene (using the same method as in Section 3.4.2). This confirmed the notion that no genuine hits had been observed in the plate screen.

The 29 other variants were all unique and fell into one of two groups (Figure 5.18). Either sequencing worked only in the forward (M13fwd primer) or in the reverse (M13rev primer) direction. This was due to a deletion which included the respective opposite primer sequence which meant that the second sequencing primer lacked its binding site on the vector. With 22 variants, the lack of the M13fwd binding sequence was most common. In 20 of these variants,

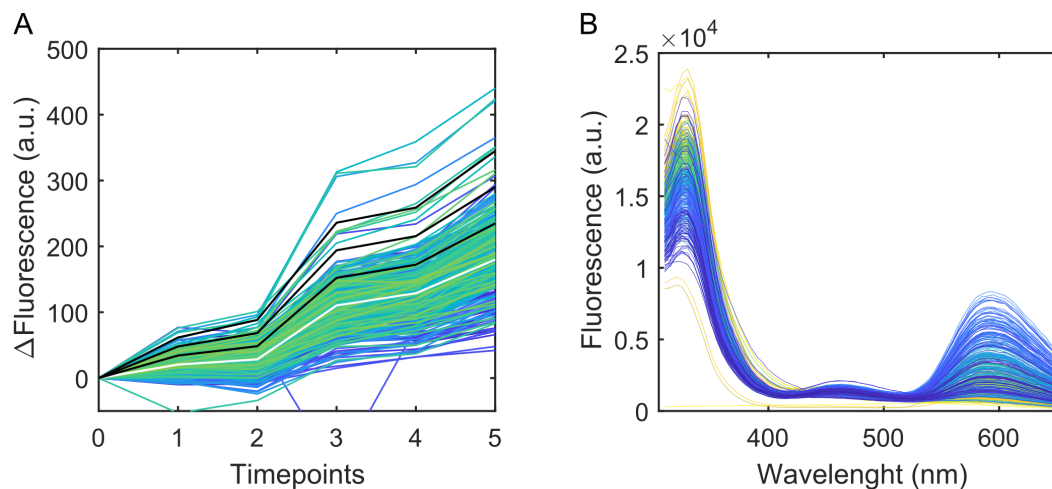


Fig. 5.17 A: Fluorescence change over time for one of three 384-well plates. The white line is the average change, the black lines represent 1, 2, and 3 standard deviations above the average. Timepoints 1: 1h, 2: 2h, 3: 16h, 4: 23h, 5: 40h. B: Fluorescence emission spectrum of the same well plate after 1 week of incubation (sealed, in the dark) showed fluorescence emission both at 335 nm and 595 nm.

the deletion started just after the *EcoRV* restriction site, had a length of 800 to 970 bp, and ended between the *f1* origin of replication and the *kanR* gene. Of the 7 variants which lacked an M13rev binding site, 6 started within a 20 bp stretch 3' of the *EcoRV* site, had a length of 210 to 500 bp, and ended just before the end of the pUC origin of replication.

Because all of the compromised vectors had deletions starting near the *EcoRV* restriction site, it is likely that these were created when the metagenomic libraries constituting the SCV library were constructed using the cloning vector pZero2. The vector pZero2 contains a C-terminal fusion of the toxic *ccdB* gene to *lacZ α* . Insertion of a DNA fragment within the multiple cloning site (MCS) disrupts the expression of the toxin. Therefore, positive recombinants can grow but not those transformed with self-ligated vector. However, it is likely that partial fragmentation of the vector occurred during the library construction. Self-ligation of fragmented vector caused the observed deletions which disrupted the function of the *ccdB* gene. The larger deletions created minimal vectors of 2.3 kbp, which retained only the essential vector elements, *i.e.* the pUC origin of replication and the resistance gene *kanR*. This compares to about 5 kbp for the average library member with insert [53, 246].

The presence of small compromised vectors in the library is highly problematic. It is likely that there is a strong positive bias amplifying them relative to larger plasmids at each transformation step. There are two factors contributing to this: 1) Colin showed that the plasmid copy number was higher for smaller inserts in vector pZero2 [84], and 2) the transformation efficiency is generally higher for smaller vectors [247]. Furthermore, small plasmids

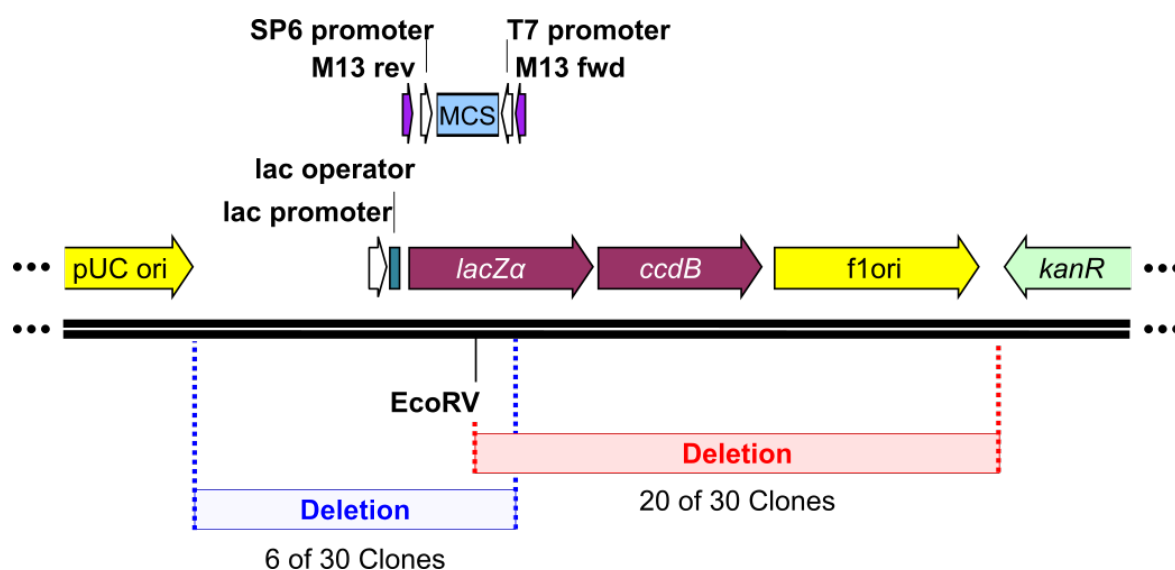


Fig. 5.18 Location of deletions observed in 29 of 30 sequenced variants. Two thirds of the variants had a deletion between the EcoRV site, used to create the metagenomic library, and the resistance gene *kanR*. About 20% of the variants had a deletion between the EcoRV site and the pUC origin of replication.

are more resistant to shear forces than larger ones [248]. Such forces can occur in ice during long-term storage of DNA [249]. Indeed, gel electrophoresis of the SCV library plasmids, which were stored at -20°C in water, showed a smear typical of DNA degradation (Appendix Figure D.6).

To assess the state of the naive, *i.e.* unsorted library, five variants were picked randomly and sequenced. All five had a deletion, showing that the quality of the library was low from the outset. As mentioned above, the SCV library is a mixture of several constituent libraries. Each of these was re-transformed and three variants sequenced to test whether or not they all had undergone degradation, see Table 5.3. All variants in library ENR-M had a deletion. One out of three variants had a deletion for libraries ENR-S, TSA, and DIR-MC/RC. With only three variants sequenced, finding any with a compromised vector was a strong indication of library degradation. Of the remaining libraries, none of the variants had a deletion. Based on these results, two new library mixtures were prepared using the original DNA samples. One was prepared as previously using all libraries (new SCV) and one excluding all libraries which had compromised variants. Thus, a new library (small SCV) with an estimated 3.5×10^5 members was created. Both mixtures were transformed again and five variants were sequenced per sample. In the new SCV library, three out of five variants contained a deletion. In the small SCV library, four out of five variants contained an insert and did not

show a deletion. Therefore, both preparing a fresh library mixture and excluding the more compromised sub-libraries improved the quality of the starting library.

Table 5.3 Listed is the number of variants with a vector deletion and no insert in the different metagenomic libraries prior to sorting. The SCV libraries are mixtures of all the other libraries. Old SCV was used in all the previous experiments, the new SCV and small SCV are fresh mixtures.

Library	Library Size /10 ⁴	Variants with Deletion	Used for small SCV
old SCV	125.3	5/5	
ENR-M	2.3	3/3	
ENR-S	3.5	1/3	
ENR-G	2.5	0/3	✓
ENR-L	3.0	0/3	✓
DIR-L	8.0	0/3	✓
SEM	8.0	0/3	✓
TSA	4.5	1/3	
DIR-MC/RC	80.0	1/3	
CR2	13.5	0/3	✓
new SCV	125.3	3/5	
small SCV	35.0	1/5	

5.3.3 Droplet screening using the small SCV library reduced false positive rate

The functional screening was repeated using the small SCV library because of its better quality. The average droplet occupancy λ was lowered to 0.1 to further reduce carry-over of non-selected cells. A total of 4.2×10^6 droplets were sorted and 158 collected yielding an apparent hit rate of 3.8×10^{-4} , slightly higher than on day 1 of the previous sorting campaign. This time, 19 of the collected events showed a fluorescence above 3 V, where previously no events had been observed. This may indicate an increased proportion of active library members in the smaller SCV library. Upon re-transformation about 5,000 colonies were obtained. Taking into account the smaller total number of droplets screened, this is 5× reduced compared to the previous sorting campaign. Therefore, an apparent reduction in false positives was achieved. Using the same calculations as previously, the apparent sorting error-rate was reduced from 10^{-2} to 10^{-3} . However, overall the number of colonies was still about 10× above the expectation. Sequencing of five randomly selected variants showed that four out of five

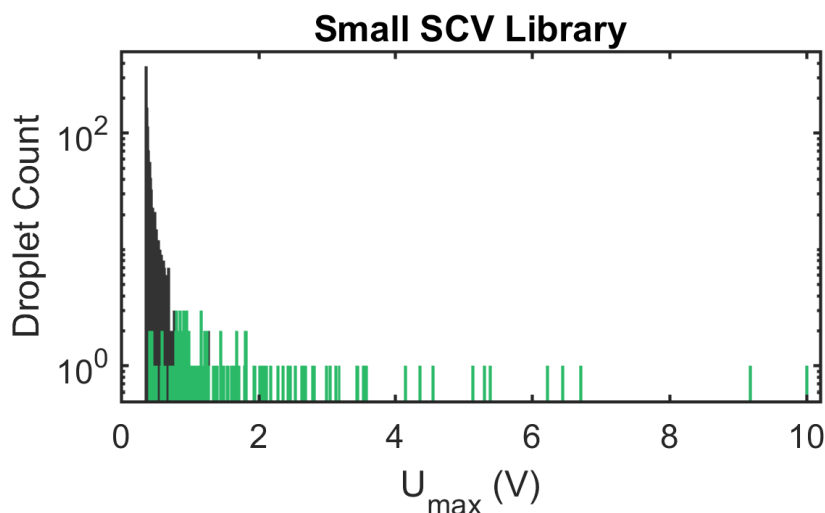


Fig. 5.19 The sorting histogram of the small SCV library (3.5×10^5 members) showed fluorescent events over the entire range of detection. There were 19 events with a signal above 3 V, higher than any previously observed signal, indicating that this library a higher proportion of highly active library members.

had a deletion. Thus, it was likely that most of the obtained colonies were again false positives and that re-screening was required.

5.3.4 Re-Screening using absorbance as primary assay readout

In the previous re-screening campaign the fluorescence change over time was followed. As discussed earlier, fluorescence emission is sensitive to small variations in the local environment of the fluorophore and thus affected by differences in cell growth and lysis efficiency. Therefore, the change in absorbance was monitored this time, as absorbance is not affected by such small variations and should thus be a more robust readout of the turnover of substrate. Differences in cell growth and lysis would be expected to affect only the absorbance offset. The assay was performed in 384-well plate format, which allowed for a longer pathlength (9.2 mm) compared to 96-well plates (2.3 mm), thus rendering the assay more sensitive by enhancing the absorbance signal by 4 fold.

As a control, HG3.17 and N20 were expressed in *E. coli* BL21-Gold and the $OD_{600\text{nm}}$ of the suspensions equalised prior to cell lysis. Each reaction was started by adding 40 μL buffer (20 mM NaPi pH 7, 50 mM NaCl), 10 μL cell lysate and 50 μL of 2 mM substrate **6a** in buffer. The reaction was monitored at 350 nm, see Figure 5.20. Over a period of three hours the signal difference between the positive and negative control reached 0.9 a.u. and doubled to 1.8 a.u. after overnight incubation in dark conditions. As expected, exposure to

UV-light at 365 nm caused appearance of the absorbance peak at 500 nm and, if excited at this wavelength, fluorescence emission at 600 nm. The absorbance at 500 nm was strong enough to see an orange hue in the positive control under daylight conditions.

These conditions were used to re-screen colonies recovered from the droplet screening of the small SCV library. Figure 5.21 shows the results for one 96-well plate³. Most reactions plateaued within 0.5 h; the last ones plateaued after 2.5 h. Assuming that most variants were false positives and that the signal had a normal distribution, the average and standard deviation derived from all wells can be used to identify outliers (2% expected at two standard deviations above the mean, and 0.1% at three standard deviations above the mean). After 3 h, the average absorbance $A_{350\text{nm}}$ was 0.50 ± 0.08 a.u. and the average change in absorbance $\Delta A_{350\text{nm}}$ was 0.11 ± 0.09 a.u.. The variants in wells A02, B06, C06 and D06 reached a signal more than two standard deviations above average. These four also had the largest relative signal change. One variant stood out as having an initial rate v_i three standard deviations above average (F12, Figure 5.22). Unfortunately, when these five variants were sequenced four had a “type I” deletion (lacking the *lacZ*, *ccdB* and *f1ori* genes) and one had a “type II” deletion (lacking the M13rev primer), grey cells were controls.

5.3.5 Combining absorbance and sequencing data yields a potential hit

Considering that large vectors are likely to be counter-selected during DNA recovery due to biological bias, any large-insert vector identified may indicate it was actively enriched during the droplet screening step. Therefore, to complement the functional assay above, the whole plate was sequenced using the standard M13Rev primer. Figure 5.23 shows a histogram of the results. It was found that 74% of the variants had a deletion, 11% inserts had inserts with fewer than 1 kbp and 15% had inserts with more than 1 kbp.

Every insert was analysed using Blastx to identify open reading frames (ORFs) with homologous sequences in the NCBI non-redundant protein database. The Blastx hits were used for putative annotation of the identified ORFs. Only six variants contained complete ORFs (Table 5.4). Assuming each of these had been selected in the droplet screening, this would correspond to a hit rate of 7%, which matches with the expected range based on the number of droplets collected and colonies recovered. Each predicted protein sequence was submitted to the Pfam database to identify the protein domains and the family they belonged to.

One ORF was predicted to encode for a transcriptional regulator, two for transporter-related proteins, one for a helicase and one for an enzyme related to kanamycin resistance. These kind of hits are typical examples of “you get what you screen for”. They can give the

³Cells were grown and lysed in deep 96-well plates and the lysates transferred to 384-well plate for the absorbance assay.

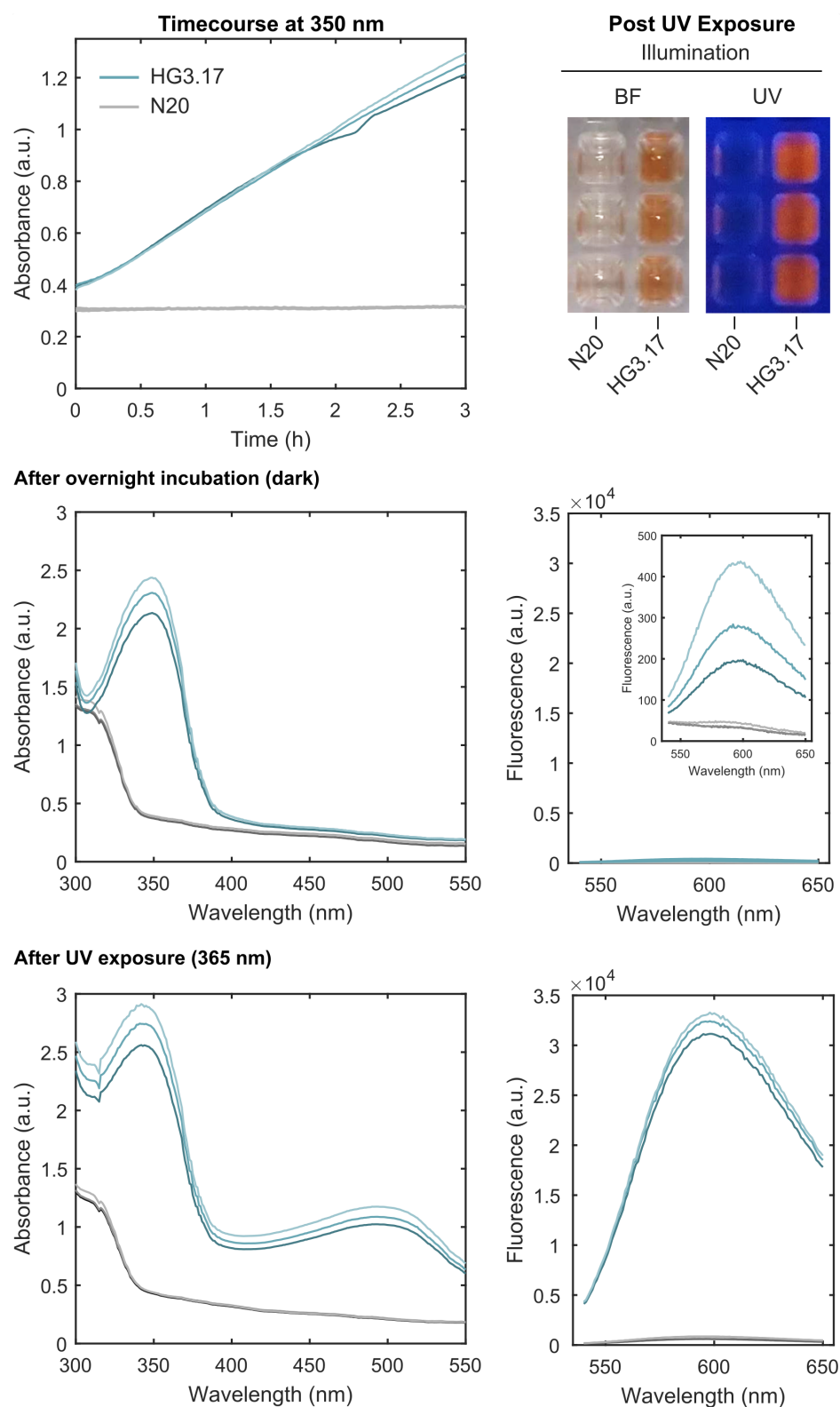


Fig. 5.20 Cell lysate controls for re-screening in 384-well plates. Triplicates of 100 μ L reactions with 1 mM **6a** in 10 \times diluted lysates of cells expressing HG3.17 or N20. The reactions were followed at 350 nm and after 3 h moved to the dark and incubated overnight (ca. 20 h total incubation time), after which they were exposed to light of 365 nm for 1 min ($\sim 100 \mu\text{Wcm}^{-2}$). Buffer: 20 mM NaPi pH 7, 50 mM NaCl, 0.1 \times BugBuster (MerckMillipore). BF: bright-field, UV: 365 nm.

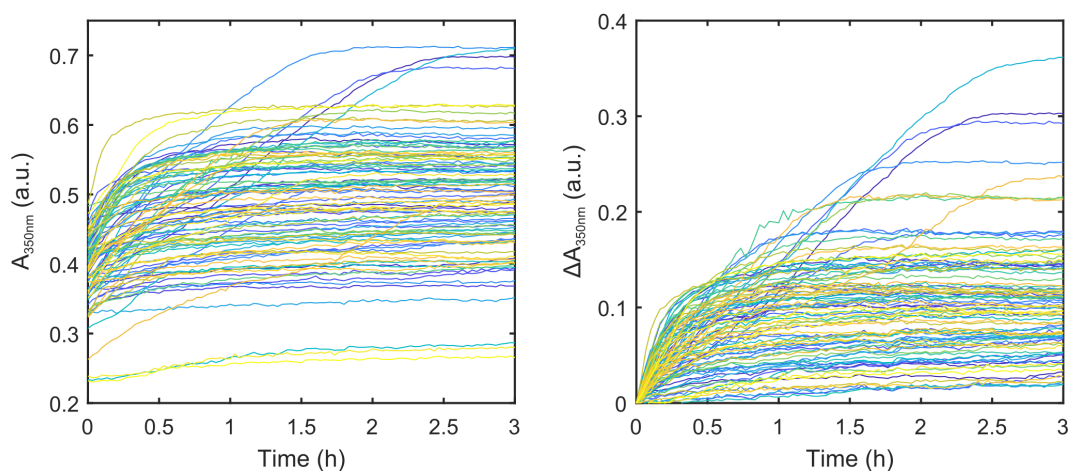


Fig. 5.21 Re-screening of the small SCV library post droplet sorting following absorbance at 350 nm using 10 fold diluted cell lysates. Conditions: 1 mM substrate **6a**, 20 mM NaPi pH 7, 50 mM NaCl, 0.1× BugBuster (MerckMillipore).

Table 5.4 Sequenced variants with inserts containing complete ORFs.

Well	Length	ORF	Protein	Pfam (Family size)
A07	2,323	1	DEAD/DEAH box helicase	DEAD (133k) Helicase_C (191k)
A08	2,418	1	PTS ascorbate transporter, IIC	EIIC-GAT (2k)
		2	PTS ascorbate transporter, IIB	PTS_IIB (10k)
A10	1,222	1	Kanamycin-modifying enzyme	Acetyltransf_1 (120k)
D12	1,136	1	Transcriptional Regulator	HTH_1 (140k) LysR_substrate (139k)
G06	1,452	1	Class IV adenylyl cyclase	CYTH (6k)
H03	1,718	1	Methyltransferase	Methyltransf_25 (50k)

variant a growth advantage (kanamycin-resistance is the selection marker for vector pZero2) or increase the background activity in the cytosol by affecting gene expression. While they thus often represent genuinely enriched sequences, it is usually based on a property different from the desired one (catalysis). Discounting these variants, two interesting hits remain: a class IV adenylyl cyclase and a methyltransferase.

In Figure 5.24 the absorbance assay data was combined with this information. It became evident that the variation in both reaction rate and total absorbance change was larger for plasmids without inserts than those containing inserts, further underlining that these are problematic during library screening. Importantly, variant G06 was found to be the most active of all variants containing a DNA insert. Taking the average (0.10 a.u.) and standard

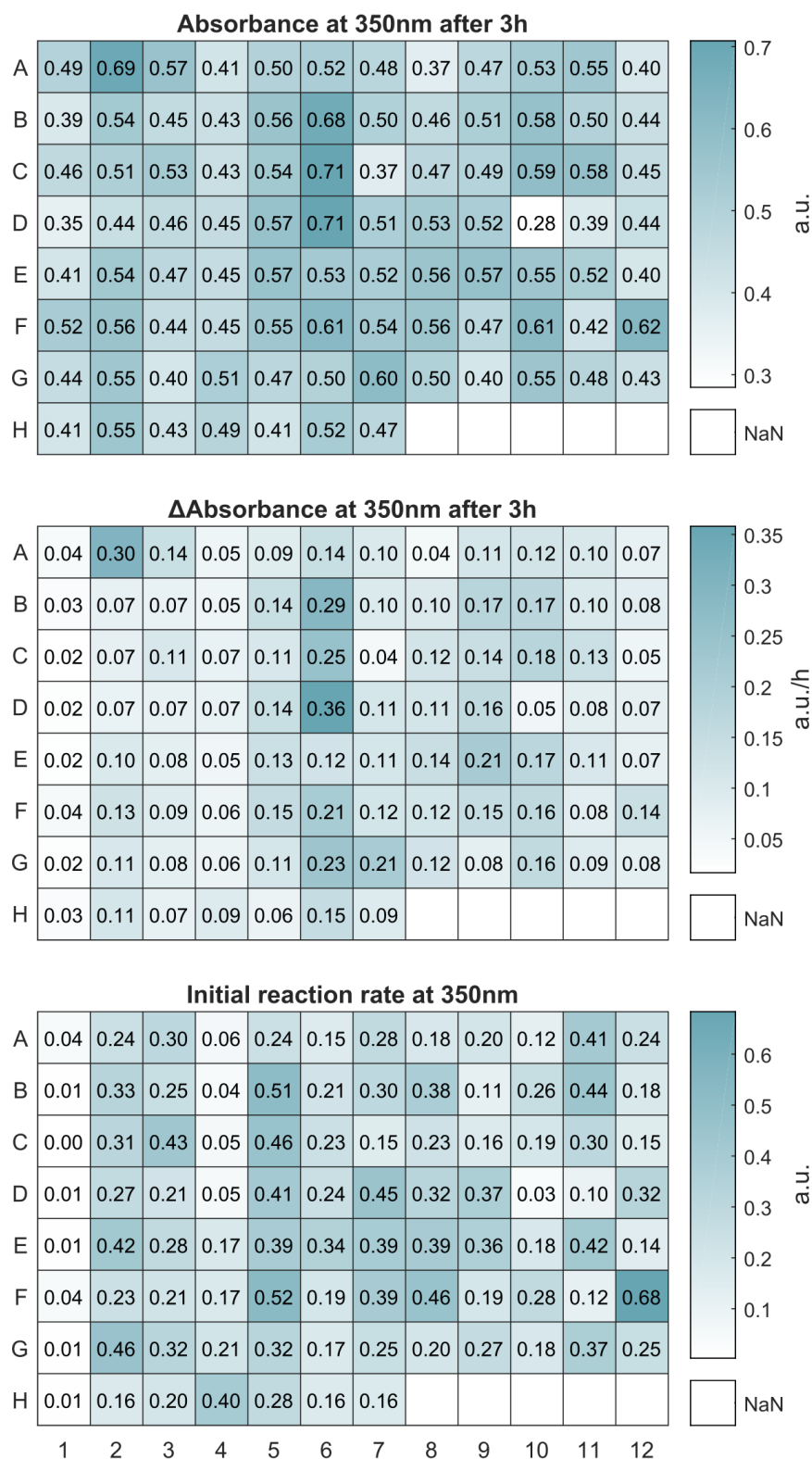


Fig. 5.22 Absorbance after 3 h and initial reaction rates observed during the re-screening of the small SCV library post droplet sorting following absorbance at 350 nm using 10 fold diluted cell lysates. Conditions: 1 mM substrate **6a**, 20 mM NaPi pH 7, 50 mM NaCl, 0.1× BugBuster (MerckMillipore).

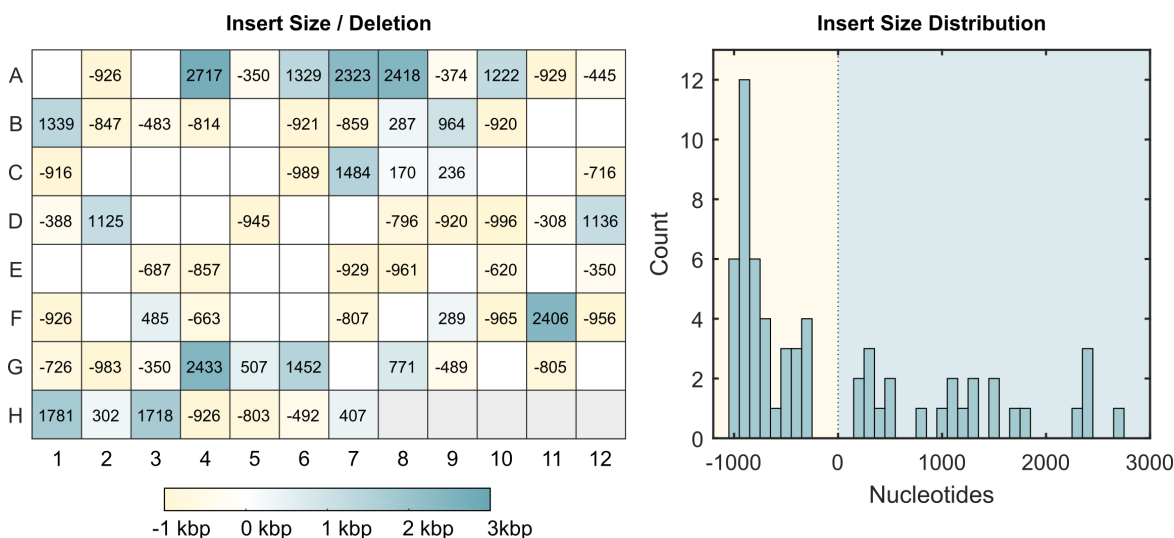


Fig. 5.23 Insert size distribution obtained by sequencing two 96-well plates of variants selected post-droplet sorting. The majority of the variants (74%) had a deletion and 1 in 10 had an insert larger than 1 kbp. Blank wells were missing the sequencing primer (“type II” deletion)

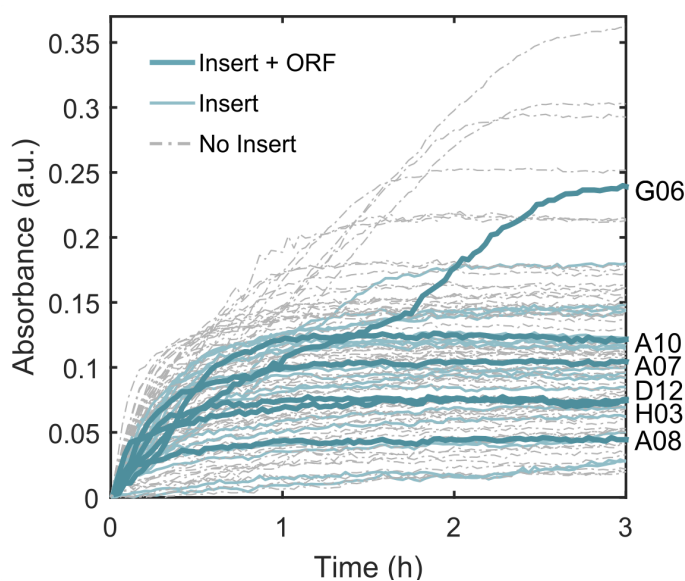


Fig. 5.24 Combining the absorbance assay data from Figure 5.21 with sequencing information reveals that well G06 harboured the most active variant with an ORF-containing insert.

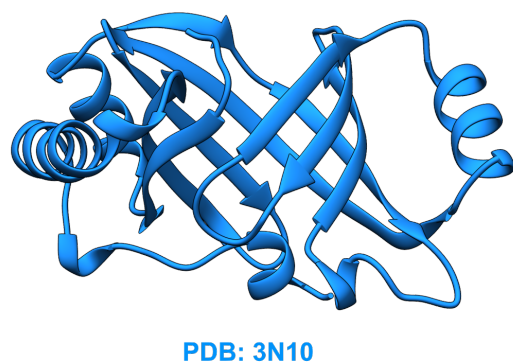
deviation of only the variants containing inserts (0.05 a.u.), G06 is close to three standard deviations above the mean (0.23 a.u.).

This variant encodes for a predicted class IV adenylyl cyclase (AC). ACs catalyse the synthesis of cyclic adenosine 3',5'-monophosphate (cAMP) from adenosine triphosphate (ATP).

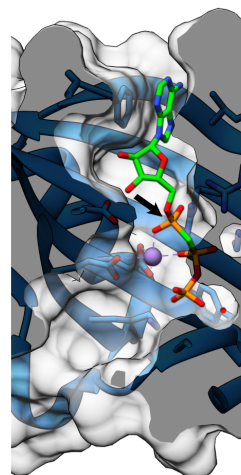
These enzymes are members of a small protein family (CYTH) and their fold is an antiparallel β -barrel with a central cavity [250]. Gallagher *et al.* resolved the structure of *Yersinia pestis* AC bound to substrate analogue α,β -methylene ATP (PDB: 3N0Y) and the product cAMP (PDB: 3N10) [251], see Figure 5.25. The catalytic mechanism is likely to involve two divalent metal ions (Mn^{2+} or Mg^{2+}) bound by conserved glutamate and aspartate residues. The metal ions coordinate the phosphate groups of the substrate and facilitate nucleophilic attack on the α -phosphate by the 3'OH. The crystal structures reveal a central substrate tunnel penetrating the centre of the β -barrel. Notably, one half of the barrel is lined with acidic residues, coordinating the metal cations, while the other side is lined with arginine and lysine residues.

The structure of variant G06 was modelled using the Phyre II web portal, which predicts protein structures based on homology with experimentally solved structures [252]. The G06 model was based on the PDB entry 2EEN of a protein of unknown function from *Pyrococcus horikoshii*, a hyperthermophilic archaeon (34% sequence identity, 48% sequence similarity). Alignment of the model with the crystal structure of the *Yersinia pestis* AC indicated conservation of the acidic amino acids that coordinate the metal cations. The top entrance to the substrate tunnel in G06 had four additional acidic residues with each two in close proximity to each other. This proximity could perturb their pK_a upwards making them efficient general bases, potentially providing catalysis on the protein surface. In the centre of the tunnel is a small hydrophobic patch formed by Tyr38, Ala52, Val110, Ser174 and Tyr175 with two lysines (Lys8, Lys 76) in close proximity – in this case possibly perturbing their pK_a downwards making them better general bases. This latter location may be a more likely environment capable of catalysing the Kemp elimination. It is reminiscent of a binding pocket in bovine serum albumin, within which a catalytic lysine promotes the Kemp elimination [193, 203].

Taking these considerations together, the evidence supports that variant G06 is a potential hit. The next steps would be to re-clone the gene for protein over-expression and purification to confirm Kemp eliminase activity of this enzyme in the absence of cell lysate. As mentioned in Chapter 4, the only comparable screening campaign was performed by Kheronsky *et al.* [205], who screened the *E. coli* ASKA library (~ 4000 variants) and found two Kemp eliminases (a hit rate of 5×10^{-4}). This indicates that G06 would have been unlikely to be found in a naive pool of fewer than 100 variants without prior enrichment by droplet screening.

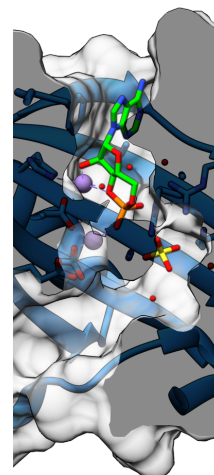
***Yersinia pestis* class IV adenylyl cyclase**

PDB: 3N0Y

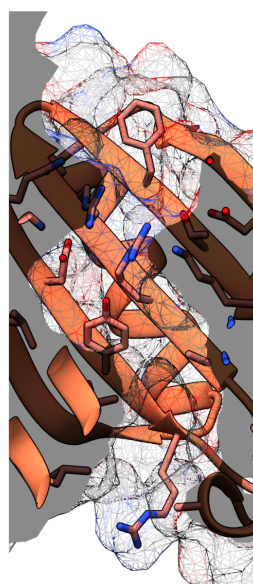
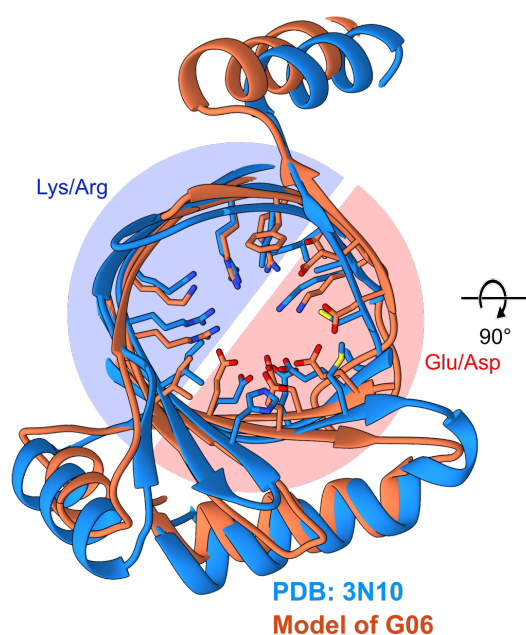


APC

PDB: 3N10



cAMP

Model of G06

180°

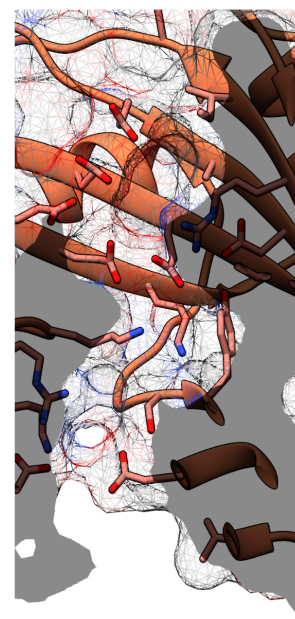


Fig. 5.25 Structure of *Yersinia pestis* class IV adenylyl cyclase bound to substrate analogue α,β -methylene ATP (APC, PDB: 3N0Y) and to product (cAMP, PDB: 3N10), respectively [251]. A model of the structure of G06, a predicted adenylyl cyclase, was generated using Phyre II [252], which revealed conservation of critical catalytic residues. Catalysis of the Kemp elimination could occur in the centre of the tunnel where two lysine residues are in close proximity near a hydrophobic patch.

5.4 Conclusions

In order to find a suitable Kemp substrate for functional metagenomic screening in droplets, several new substrates were synthesised. One of these substrates, **6a**, was found to become fluorescent after conversion to the Kemp product and brief irradiation at 365 nm. Fluorescence measurements are generally far more sensitive than absorbance measurements [253]. Only one other fluorogenic Kemp substrate has been reported to date [229]. However, this coumarin-based substrate was not used in the study of enzymology, but rather to quantify serum albumin in blood for diagnostic purposes. Yet, the potential to detect lower quantities of Kemp reaction product offers new possibilities in enzyme engineering. For example, most Kemp eliminases have a Michaelis constant K_m in the milli-molar range, whereas the typical K_m of natural enzymes is in the micro-molar range [158]. This can be attributed to the fact, that absorbance-based screening campaigns are limited to substrate concentrations in the milli-molar range because of limited sensitivity. A fluorogenic substrate could overcome this limitation in future evolutionary campaigns. The substrate **6a** seems particularly suited for this, because of its structural similarity to **2a** it would be more comparable to existing literature as compared to the coumarin-based substrate by Sakamoto *et al.*

It was possible to show that **6a** is converted to the expected Kemp reaction product in the first step. The structure of the fluorophore generated upon irradiation with UV light has eluded full characterisation due to low solubility in H_2O . The available data indicates only a small change in structure, *e.g.* the rotation around the azide-carbon bond, which possibly allows the formation of an exciplex upon excitation. Crystallography of the final compound would yield strong evidence for its molecular structure. However, dynamic in-solution behaviour may play an important role in the mechanism of fluorescence. Future studies should therefore include 1H -NMR of a dilution series of the final compound in D_2O to study the complexation behaviour, but will require very long integration times.

For the purpose of this study, the focus remained on establishing a functional metagenomic screening platform. Therefore, a second droplet assay was developed. In this assay, the droplet diameter could be reduced to 15 μm diameter (a volume of 1.8 pL). This diluted the cytosol of a single cell 100 \times less than in the absorbance assay thus likely lowering the limit of detection. Again, the selectivity of the assay was shown by enriching the positive control HG3.17 over the negative control.

Using this substrate for two metagenomic screening campaigns, it was possible to detect fluorescent events above background. The apparent hit-rate (2.7×10^{-4}) was comparable to what was observed in the esterase screening campaigns (4.5×10^{-5}). The fluorescence signal of about 10% of these apparent hits exceeded the highest signal obtained for HG3.17.

Together these observations based on screening a large number of library variants indicated that efficient Kemp eliminases are more common than assumed until now.

However, a high proportion of false positives was obtained after recovery of the DNA from the collected droplets. Re-screening in plates based on the readout of fluorescence or absorbance alone was not sensitive enough to recover the hits. Sequencing of a whole plate revealed that vectors with deletions had a higher variation in signal compared to vectors containing inserts, thus obscuring the functional readout. Using the sequencing data it was possible to identify a promising lead, variant G06, which may be the first Kemp eliminase identified in a metagenomic screening campaign.

This work highlighted the limitations placed on functional metagenomic screening by poor library quality. Construction of a new high-quality metagenomic library would a) greatly reduce the number of false positives and b) simplify the re-screening procedure due to reduced variation of the background activity. This may then yield a much greater number of potential Kemp eliminase hits. The screening method, for indeed any enzymatic activity, could be further complemented by a deep-sequenced metagenomic library in which the frequency of each ORF is known - this would allow to compute the likelihood of finding each ORF by chance and give more confidence in it having been genuinely enriched post-sorting, even without the need for sequencing the whole library output.

Chapter 6

Conclusions

In this thesis, I set out to meet three goals: building a FADS, establishing a functional metagenomic screen for esterases using the FADS, and establishing a functional metagenomic screen for Kemp eliminases using AADS.

Chapter 2 described how I implemented a state-of-the-art FADS, which contributed to the success of several droplet sorting campaigns, including a collaboration on directed evolution of a sulfatase using *E. coli* autodisplay and directed evolution of a protease using IVTT. Traditionally, when screening enzyme libraries, single variants need to be separated physically on a macroscopic level, *e.g.* into different wells, to be able to measure their relative activities. The screening of enzyme libraries in microfluidic droplets differs, because the library variants remain pooled in a water-in-oil suspension. Conceptually, it therefore resembles the screening for binding proteins, *e.g.* antibodies, in which successive rounds of screening enrich for improved binders [254, 255]. While the utility of a new system can be verified by the enrichment of a positive control diluted with a negative control [254], the ultimate test for the system is the screening of a library of variants [255]. The FADS reported here passed this test for different enzyme assays and libraries.

In Chapter 3, I used the FADS to set up an esterase assay and reported the first functional metagenomic screen for esterases using droplet microfluidics. Twelve plasmids were isolated from a million-membered metagenomic library encoding for thirteen novel esterase genes, which were confirmed to be efficient esterases. The absolute hit rate was 10^{-5} . In general, hit rates vary strongly between different library formats due to different insert sizes and environmental sources, for esterases they have been reported from as low as 10^{-7} to as high as 10^{-4} [31]. One previous study, which used plasmids with similar insert sizes as here, had the same hit rate as the droplet microfluidic assay reported here. However, in this study over $4\times$ as many esterases were isolated, because $4\times$ as many clones were screened. Of the thirteen new genes, ten were from the large α/β -hydrolase super-family of proteins (close to

500,000 sequences on the Pfam database). Two of the hits, N20 and RR11ORF2 were from the DUF3089 family, which contains only 500 sequences. A sequence belonging to a domain of unknown function family would not have been isolated by a sequence-based metagenomic screening, especially not from such a small one. Notably, these two enzymes were the most active esterases reported here. Several other hits were from very small protein families, such as N13 (Lipase_Bact_N, which with 200 sequences is too small and too distantly related to even assign it to a super-family). The high proportion of hits from small protein families indicates that the number of sequences present in a protein family is not a good predictor of the abundance of these sequences in the environment nor the level of enzymatic activity of their gene products. Again, a sequence-based methods would have been biased towards larger families suspecting to find better catalysts in extended sequence clusters. In contrast, functional metagenomic screening provided access to thinly populated sequence-space containing efficient esterases without the need for *a priori* knowledge.

As mentioned, hit rates can vary strongly between libraries. Here, the cow rumen sub-library had a hit rate 6× higher than the other 9 sub-libraries combined (40% of the hits from 10% of the screened clones). In another large-scale screening study for esterases, 40% of the hits were found in 4 of the 17 sub-libraries (5% of the screened clones) [127]. These findings can be used to guide future research. For example, instead of pooling many libraries to obtain one large library, the libraries could be pre-screened separately. Using droplet microfluidics, even very low apparent hit rates can be measured quickly. If the FADS is used in an “analytical” mode without sorting, the droplet rate can be increased, because the limiting factor in FADS are the dynamics of sorting, not determining the fluorescence signal of each droplet. Thus, it should be feasible to determine apparent hit rates of ten libraries in a day’s work at low reagent cost. Then, the libraries with the highest apparent hit rates could be pooled and screened at very high coverage to ensure recovery of every clone (a library with 10⁵ members would have been oversampled 100× under the conditions used in Chapter 3). This method could also be used with with cosmid- or fosmid-based libraries. Cosmids and fosmids tend to have higher hit rates, but low copy numbers (>50 per cell). Isolation of the low-copy cosmids or fosmids from single droplet has not been published to date. However, under a pre-screening regime the isolation of the DNA would not be necessary. Several libraries could be pre-screened and the one with the highest hit rates than re-screened using traditional plate methods. This method may also reduce bias and the contribution of false positives from low-quality metagenomic libraries.

In Chapters 4 and 5, I established droplet microfluidic assays for the Kemp elimination based on absorbance using substrate 5-nitro-1,2-benzisoxazole (**2a**) and the newly discovered fluorogenic substrate 5-azido-1,2-benzisoxazole (**6b**), respectively. Substrate **6b** is struc-

turally very similar to the substrate **2b**, which has been most widely used to date because it can be detected most sensitively in a colorimetric assay.

Both Kemp eliminase assays were able to enrich a positive control for the Kemp elimination over a negative control. The absorbance-based assay was not sensitive enough for functional metagenomics, but its utility was shown in the screening of several mutant libraries of the Kemp eliminase HG3.17. The fluorescence-based assay was able to detect Kemp eliminase activity in droplets. The apparent hit rate suggested that promiscuous enzymes capable of catalysing this reaction are commonplace, but the recovery of hits was impeded by the low quality of the metagenomic library SCV (discussed in more detail further below). A combination of functional and sequence-based re-screening lead to identification of a possible lead - library member 4G06, which is predicted to encode for a class IV adenylyl cyclase. This initial lead indicates that screening of good quality metagenomic libraries in the future should enable the isolation of promiscuous Kemp eliminases based on the assay developed here.

In general, substrate **6b** should enable more sensitive product detection at lower substrate concentration, therefore also be suited for the evolution of Kemp eliminases towards lower Michaelis constants and thus higher activities than previously possible. In the case of HG3.17, which fddid was evolved from HG3, k_{cat} improved from 3 s^{-1} to 700 s^{-1} , whereas there was no change in K_{m} with 2.4 and 3.0 mM respectively [16]. This indicates, that the screening was performed in saturating conditions, *i.e.* the observed reaction rate was close to v_{max} , and therefore a change in K_{m} of the enzyme did not significantly influence the observed initial rate. However, if the substrate concentration $[S] \ll K_{\text{m}}$, then the Michaelis-Menten equation (Equation 3.1) simplifies to: $v_i = (k_{\text{cat}}/K_{\text{m}})[S][E_0]$. If screened in such conditions, both changes in k_{cat} and K_{m} would influence the initial rate and therefore the selection pressure on K_{m} would be increased. A fluorogenic substrate should enable screening using substrate concentrations in the 10 to 50 μM range (as in the esterase screening), which is 10 \times below the concentrations used in the directed evolution of HG3.17 [16]. A directed evolution campaign using this substrate in microfluidic droplets, should additionally benefit from ultrahigh-throughput, enabling the screening of different kinds of libraries, *e.g.* insertion and deletion libraries, which are currently not routinely used in enzyme engineering because of the large fraction of inactive variants usually observed [211].

Instead of HG3.17, the hit N7 (Chapter 3) found to be a promiscuous catalyst of the Kemp elimination could be used to start from a lower Kemp eliminase activity. It has been observed that the improvement during the initial rounds of directed evolution for a promiscuous activity are large and start to diminish [256, 257]. This was also observed for HG3.17 [16]. It

may therefore be easier to use N7 as a starting point to verify **6b** as a useful substrate for directed evolution, and compare its evolutionary trajectory to HG3.17 in a second step.

Analysis of sources of error and bias in droplet sorting A high number of false positives was observed in both the esterase and the Kemp eliminase screening campaigns. To improve future screening outcomes, it is useful to analyse the possible causes of this observation. There are two separate aspects to this. The first one is the ratio of positives to negatives. It is possible that false positives were actively collected because the sorting thresholds were close to background levels. However, the second aspect is the total number of colonies recovered per droplet. In control experiments this was in general between two and four, therefore in any of the metagenomic screens about 10^3 colonies would have been expected, but $10\times$ more were obtained. The main question is therefore, where these excess colonies came from.

Figure 6.1 summarises possible sources of these colonies. In a system free of background, error, and bias, the isolation of a library variant would purely depend on the sensitivity of the droplet assay and the ability to amplify the collected DNA by recovery and transformation of fresh cells. However, in a real droplet system there are three possible sources of error:

- co-encapsulation of positive clones with negative ones in droplets due to the Poisson distribution;
- erroneous droplet collection due to random fluctuations of the droplet flow during FADS; and
- active droplet collection of negative clones due to phenotypic variation between cells of identical genotype.

These three types of error allow negative variants to circumvent the screening barriers, creating false positives. The first two points are purely technological issues and their contribution to false positives can be calculated.

Co-encapsulation is only a minor contributor to the fraction of false positives. If positive and negative variants are considered two independent cell populations, the chance of co-encapsulation can be calculated as the product of the respective Poisson probabilities [60]. Using the average droplet occupancy λ of 0.35, the number of droplets sorted, and the number of hits collected in esterase sorting campaign I, the number of “passenger” cells can be calculated to have been 83. If every collected plasmid had equal recovery and transformation efficiencies, only 26% of recovered colonies would have been false positives¹.

¹These calculations apply to random co-encapsulation due to Poisson statistics. Non-random co-encapsulation happens if cells stick together. *E. coli* forms biofilms if grown at high densities under non-ideal

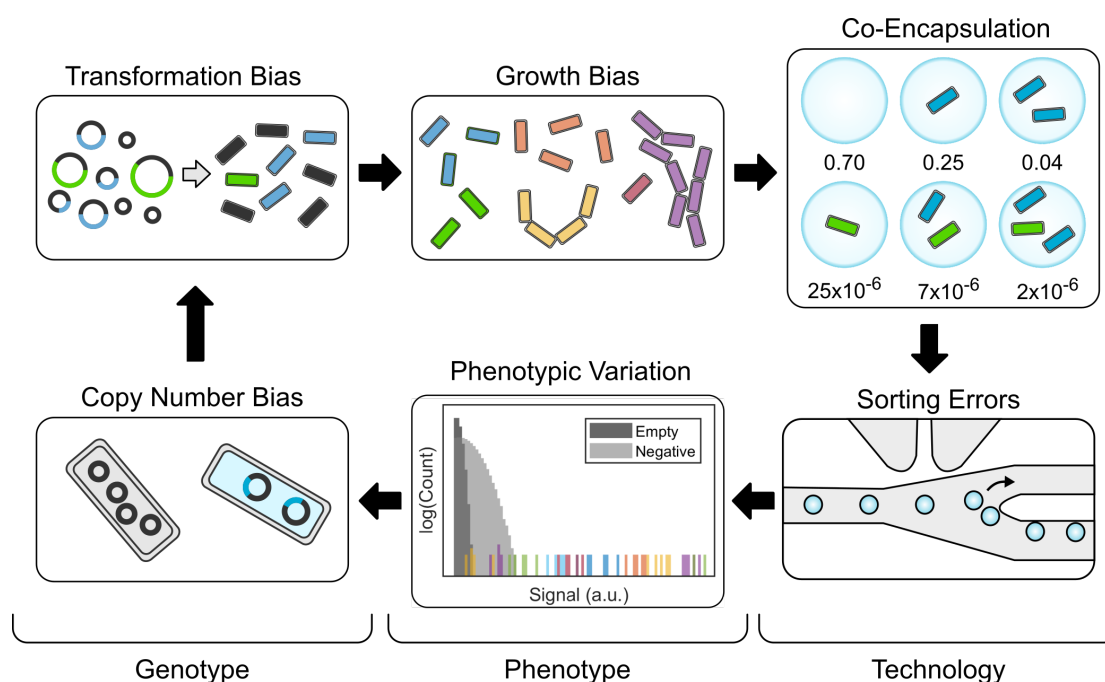


Fig. 6.1 Sources of error and bias in functional metagenomic screening that influence the fraction of false positives. The contribution of each of these factors is discussed in the main text.

The second technological source of error is due to random collection of droplets, for example due to a piece of dust passing through the chip or two droplets being sorted together. During the sorting campaigns, the occurrence of such events was monitored by high-speed video analysis. They were not observed frequently, but some such events may have been missed. The only published account of the error rate in FADS found it to be less than 10^{-4} [60]. An error rate of 10^{-4} would therefore be a worst case assumption²; and it would have allowed an additional 549 cells to enter the collection channel, setting the fraction of false positives to 73%, *i.e.* 1 in 4 colonies should have been a genuine positive.

This is still two orders above the hit rate actually observed during re-screening on tributyrin plates. Therefore, the contribution of biological errors and biases to the reduction in hit frequency exceeds the contribution of pure technology by at least two orders of magnitude.

Phenotypic variation is one source of biological error and affects both the false positive and false negative rate. It is the general phenomenon that organisms of the same genotype display different phenotypes, in the way that even monozygotic twins differ from each other.

conditions [258]. If this happens, cells can be observed to form a biofilm even within a microfluidic device during droplet generation (Appendix Figure D.7). This effect can be avoided by growing the cells at a controlled temperature and humidity; thorough washing steps; and filtering through a 5 μm filter.

²At sorting rates up to 10^3 an error-rate above 10^{-4} would have been easily observed by video analysis.

On a cellular level, this can be observed in the stochastic expression of genes which is rooted in the statistical mechanics of chemical systems at very low concentrations [259, 260]. That is, the amount of protein expressed can vary strongly between genetically identical cells. The standard deviation from the mean expression level has been found to be anywhere between 10 and 100% [260]. As mentioned, this may cause some false positive clones to be actively collected, for example due to the random up-regulation of an esterase encoded in the *E. coli* genome. But even if the majority of the collected droplets would have been false positives, this effect does not explain the high number of colonies recovered from the droplet experiments.

Therefore, the only remaining explanation is recovery bias in favour of false positives, which can be due to:

- a variation of growth rate between library members;
- a negative correlation between plasmid copy number and plasmid size (affecting the recovery efficiency from droplets); and
- a negative correlation between transformation efficiency and plasmid size.

The growth rate can vary between library members, if *e.g.* an expressed gene is toxic to the host or gives it a fitness advantage. These effects are unpredictable in metagenomic libraries and were mitigated by growing the library on culture plates rather than in solution. The other two factors were discussed in Section 5.3.2, and are likely to have influenced the sorting outcome the most. They play a role only if there is a strong variation in the size of the plasmids making up the library which is screened, as was found to be the case for the SCV library due to the presence of vectors without inserts. These insights lead to three recommendations for future screening campaigns.

Recommendations to improve functional metagenomic droplet sorting Based on the analysis above, the best chance to improve the outcome of functional metagenomic screening in droplets is to obtain or construct a better quality metagenomic library and possible modification of the workflow.

- If constructing a new metagenomic library, great care should be taken to avoid fragmentation or over-digestion of the vector. Furthermore, a cryoprotectant such as glycerol should be added to samples used for long-term storage to avoid freezing and thus shearing of the DNA [249].
- A second way to reduce bias, could be to avoid cell-lysis and re-transformation. The cells could be encapsulated in minimal medium without lysis agent to allow limited

growth of the cells in droplets. Spontaneous lysis of one or more cells in a droplet would be sufficient to detect enzymatic activity. The sister cells could then be re-grown directly without the need for re-transformation. The success of this strategy would depend on the relative influence of transformation and growth bias.

- Finally, the bias could be reduced by using libraries with larger inserts, such as fosmids with around 40 kbp. Good quality fosmid libraries are freely available from the repositories of specialised research groups [37]. Assuming, for example, a standard deviation of 1 kbp of the insert size, it is evident that the physico-chemical differences between a 39 kbp and a 40 kbp construct (a 2% difference) are smaller than between a 3 and a 4 kbp plasmid (a 25% difference). It is also easier to separate vectors with and insert from vectors without inserts during library construction. The limitation of fosmids is their low copy number which poses a challenge to DNA recovery from single cells, but in combination with the previous point, this may be a viable option.

Together, these recommendations will reduce the frequency of false positives. This will help to exploit the full power of functional metagenomics at ultrahigh-throughputs for the isolation of a larger number of novel enzymes.

Chapter 7

Methods

7.1 Microfluidics

7.1.1 Fabrication of microfluidic devices

The microfluidic devices were fabricated following classical soft-lithography procedures using high-resolution acetate masks (custom prints by Microlithography Services Ltd) and SU-8 photoresist patterning [55]. After creating access holes using a biopsy punch (1 mm, Kai Medical), the poly(dimethyl siloxane) (PDMS, Dow Corning) micro-channels were plasma bonded to a microscope glass slide in a low-pressure oxygen plasma generator (Femto, Diener Electronics, Femto), flushed with 1% v/v trichloro(1H,1H,2H,2H-perfluorooctyl)silane (Sigma-Aldrich) in HFE-7500 (3M), and left at 65 °C for 30 min.

In the case of the absorbance device, the obtained PDMS microchannels were plasma bonded to a thin cured PDMS layer (5 g PDMS in a Petri dish, ø 9 cm), flushed with 1% v/v trichloro(1H,1H,2H,2H-perfluorooctyl)silane (Sigma-Aldrich) in HFE-7500 (3M), left at 65 °C for 30 min, and the PDMS-PDMS devices then bonded to a microscope glass slide. The optical fibres (SMA Fiber Patch Cable, ø 50 µm, 0.22 NA, Thorlabs) were tripped so that only the core and the cladding were manually inserted into the chip under a microscope. The alignment was tested by connecting one fibre to a 385 nm LED light-source (M385FP1, Thorlabs) and the other to a photo-detector (PDA100A-EC, Thorlabs). The obtained signal was maximised by adjusting the position of the fibres. They were then fixed with epoxy glue and freshly mixed and degassed PDMS was inserted into the fibre channels and cured overnight at room temperature.

7.1.2 Generation and incubation of droplets

Droplets were produced using microfluidic double flow-focusing junctions. Two aqueous streams were mixed at the first junction and dispersed into the fluoruous oil HFE-7500 (3M) containing 1% w/w fluorosurfactant-008 (RAN Biotechnologies) at the second junction. The dimensions at both junctions in width by height were 15 by 16 μm and 50 by 80 μm for fluorescence and absorbance assays respectively resulting in droplet volumes of about 2 and 200 μL . The process was monitored on an inverted microscope (SP981, Brunell Microscopes) equipped with a high-speed camera (Miro eX4, Phantom Research). The microfluidic devices were operated using syringe pumps (Nemesys or Cetoni) and gas-tight syringes (SGE) which were connected to the chip *via* fine bore PTFE tubing (ID 0.38 mm, OD 1.09 mm, Smith Medical). The standard flowrates were 50 $\mu\text{L h}^{-1}$ for each aqueous and 500 $\mu\text{L h}^{-1}$ for the oil phase for the small droplets. For the large droplets they were 6 $\mu\text{L min}^{-1}$ for each aqueous and 37.5 $\mu\text{L min}^{-1}$ for the oil phase.

The generated droplets were collected into an inverted 500 μL microcentrifuge tube which was pre-filled with fluoruous oil containing the surfactant. The tube was modified by inserting tubing through access holes at the top and bottom of the microcentrifuge tube. This incubation chamber was sealed with adhesive glue (Scotch-Weld PR1500, 3M). The droplets were collected via the top tubing while the bottom tubing served as a drain. After droplet generation, a gas-tight syringe was connected to the bottom tubing in order to eject the droplets back out for downstream applications. The droplets were incubated under quiescent conditions at room temperature (*ca.* 22 $^{\circ}\text{C}$) and in the dark if a fluorogenic substrate was used.

7.1.3 Droplet leakage assays

Two populations of droplets were generated at a 1:1 ratio on a single microfluidic device comprising two separate flow-focusing devices (width 50 μm , height 80 μm at the flow-focusing junction). One population contained buffer as indicated and the other contained the buffer and the product at 1 mM. The flow rates during droplet generation were: 37.5 $\mu\text{L min}^{-1}$ and 12 $\mu\text{L min}^{-1}$ for the oil and aqueous phases respectively. For short incubation times, the droplets were collected into PTFE tubing which had been pre-filled with fluoruous oil and was directly connected to the absorbance chip. For longer incubation times the droplets were collected into an inverted reaction tube as described above. Fractions of the droplets were analysed after different time intervals as described in Section 7.1.5.

7.1.4 Fluorescence-activated droplet sorting (FADS)

The FADS and different stages of its development are described in detail in Chapter 2. In general, droplets were injected from the modified reaction tube (Section 7.1.2) into the sorting chip at $5 \mu\text{L h}^{-1}$ and spaced with plain fluoruous oil HFE7500 at $100 \mu\text{L h}^{-1}$ resulting in a sorting frequency of about 300 Hz. Depending on the quality of the droplet sample, the droplet injection rate was increased up to $15 \mu\text{L h}^{-1}$ (to a rate of 2,000 Hz) and the oil flow adjusted to create enough separation to sort single droplets.

The chip was monitored using the microscope (IX73, Olympus) light source with a long-pass filter (593nm, BrightLine Semrock) and a high speed camera (Miro eX4, Phantom Research). To measure droplet fluorescence, a laser beam (488nm, 30 mW, 85 BCD 30 Melles-Griot, attenuated with ND 1.0) was expanded 10 \times and focused onto the microfluidic channel upstream of the sorting junction *via* a dichroic mirror (495 nm, Olympus). The induced fluorescence was collected by an air objective (LUCPlanFLN 40x/0.6, Olympus), passed through a longpass filter (488nm, RazorEdge Semrock), a dichroic mirror (555nm, Thorlabs), and finally a bandpass filter (525/28nm, BrightLine Semrock) before reaching the detector. During the first metagenomic sorting campaign an avalanche photodiode (SPCM-AQRH-13, Excelitas Technologies) was used to record the signal. For all other reported experiments a photomultiplier tube (PM002, Thorlabs) was used due to greater flexibility.

In the first version of the FADS, the signal from the photodetector was split in two. On one line, it was recorded (6341, NI) and visualised using a custom LabView program. On the other, the sorting decision was taken by an ATmega328 microcontroller (Arduino Uno). The signal was either digital (APD) or analog (PMT) and analysed accordingly. In either case, if the signal exceeded a set threshold, a trigger pulse was sent to the pulse generator (TGP110, Thurlby Thandar Instruments). In the latest version of the FADS, the signal was fed to an anaolog-in pin of a Virtex-5 LX30 FPGA (PCIE-7841R, NI). The FPGA quantified the width at threshold, width at half-maxium, area, and amplitude of each droplet signal. The droplet data was streamed to a custom LabView program for visualisation and the setting of sorting thresholds. If a droplet met the sorting criteria, a trigger pulse was sent to the pulse generator.

If triggered, the pulse generator created a 5 V pulse with 500 μs width. This pulse controlled a function generator (20 MHz DDS Function Generator TG 2000, TTI) working in external gated mode, generating a 10 kHz square signal at an adjustable amplitude (typically 7 to 9 V), which was then amplified 100 times with a voltage amplifier (TREK 601c) to actuate the sorter. The electrodes were made by filling channels with salt solution (5 M NaCl) and connected to the amplifier via syringes [102].

7.1.5 Absorbance-activated droplet sorting (AADS)

Droplets were injected from the modified reaction tube (Section 7.1.2) into the sorting chip at $2\ \mu\text{L min}^{-1}$ and spaced with plain fluoruous oil HFE7500 at $20\ \mu\text{L min}^{-1}$ resulting in a sorting frequency of about 100 Hz. The absorbance signal was generated using a 385 nm LED light-source (M385FP1, Thorlabs) monitored by a photomultiplier tube (PDA100A-EC, Thorlabs). The principle sorting set-up was described by Gielen *et al.* [61]. The chip design, signal acquisition and sorting actuation was identical. The signal processing was extended to quantify the time each droplet spent within the detection area as a measure of its size. Paul Zurek extended the software on the Arduino Due to measure peak width and minimum. I extended the LabView software to analyse the signal in the equivalent way.

7.1.6 Droplet conditions for the metagenomic esterase screen

Droplets were generated, incubated, and sorted as described above. The two aqueous phases used to generate the droplets were a suspension of cells expressing the library and a solution containing substrate and lysis agents. The latter solution was composed o $20\ \mu\text{M}$ fluorescein dihexanoate (gift by Ana Torrado Agrasar), 0.4x BugBuster (BugBuster 10X Protein Extraction Reagent, Merck-Millipore), and 27 to $33\ \text{U}/\mu\text{L}$ rLysozyme (Merck-Millipore) in the buffer ($50\ \text{mM}$ TrisHCl pH 8.0, $100\ \text{mM}$ NaCl, $50\ \mu\text{g mL}^{-1}$ kanamycin, and one tablet/50ml complete EDTAfree protease inhibitor by Roche). The cell suspension was prepared as follows. Transformation of $10\ \text{ng}$ of the metagenomic library into *E. coli* (E. cloni 10G Elite, Lucigen) yielded 5×10^7 variants on Luria Bertani (LB) agar plate (containing $50\ \mu\text{g mL}^{-1}$ kanamycin) covering the library size about 50times. The bacteria were grown overnight at $37\ ^\circ\text{C}$, then incubated at room temperature (*ca.* $22\ ^\circ\text{C}$) for two days. Colonies were subsequently scraped from the plates using $3 \times 3\ \text{mL}$ liquid LB, cooled on ice, and washed three times by centrifugation at $5,000\ \text{g}$ for 3 min and resuspension in buffer. The $\text{OD}_{600\text{nm}}$ of the washed cell suspension was measured and diluted to OD 0.8 in buffer with Percoll (Sigma-Aldrich) at 25% v/v final. This suspension was expected to result in a droplet occupancy of $\lambda = 0.35$ when diluted 1:1 on chip and generation of 2 pL droplets.

7.1.7 Droplet conditions for the Kemp eliminase assays

Conditions for AADS

The two aqueous solutions for droplet generation were composed as follows. The substrate/lysis solution contained $1\ \text{mM}$ 5-nitrobenzisoazole, 0.2x BugBuster (BugBuster 10X Protein Extraction Reagent, Merck-Millipore), 27 to $33\ \text{U}/\mu\text{L}$ rLysozyme (Merck-Millipore), and 10%

v/v methanol in the buffer (20 mM TrisHCl pH 7.0, 50 mM NaCl, and 3 mM tartrazine). The cell suspension with cell expressing the metagenomic library was prepared as described in the preceding section. If the cells were *E. coli* BL21 with a pHAT based library, they were scraped on the day before droplet generation, 1 mL of undiluted cell suspension was added to 100 mL of liquid LB (with 100 $\mu\text{g mL}^{-1}$ ampicillin), and the cells allowed to rest for 1 h at 20 °C with shaking at 220 rpm. Overnight expression was then induced with 400 μM IPTG.

The next day, the cell suspension was washed three times by centrifugation at 5,000 g for 3 min and resuspension in buffer. It was diluted to OD_{600nm} of 0.008 in buffer with Percoll (Sigma-Aldrich) at 25% v/v final. This suspension was expected to result in a droplet occupancy of $\lambda = 0.35$ when diluted 1:1 on chip and generation of 200 pL droplets.

Conditions for FADS

The two aqueous solutions for droplet generation were composed as follows. The substrate/lysis solution contained 2 mM 5-azido-1,2benzoxazole, 0.4x BugBuster (BugBuster 10X Protein Extraction Reagent, Merck-Millipore), and 27 to 33 U/ μL rLysozyme (Merck-Millipore) in the buffer (20 mM sodium-phosphate pH 7.0, 50 mM NaCl). The cell suspension was prepared as described in Section 7.1.6 using the buffer specified here.

The incubation of the droplets was modified as follows. They were collected into an inverted 200 μL PCR tube modified with tubing at the top and bottom. After droplet generation, the tubing was sealed and the PCR tube was mounted to a custom made UV-exposure device controlled by an Arduino Uno. Every 15 min the droplets were mixed by a servomotor (SM-S2309S, SpringRC) rotating 90° left, waiting for 2 min, rotating 180° right, waiting for 2 min, and returning to the original orientation by rotating 90° left. The tube was then exposed for 15 s to an expanded beam of light from a 365 nm LED (M365F1, Thorlabs) which was run at 400 mA.

7.1.8 DNA Recovery from microfluidic droplets

Sorted droplets were collected into a 1.5 mL low DNA retention reaction tube (DNA LoBind, Eppendorf) prefilled with 20 μL fluoruous oil HFE7500. The same tubes and low DNA retention pipette tips (Low Retention, Axygen) were used for all subsequent steps to minimise DNA loss. Of a solution of 2 ng μL^{-1} salmon sperm DNA (Invitrogen), 100 μL was added to the collected droplets. To the oil layer, 10 μL of 1H,1H,2H,2H-perfluorooctanol (97%, Alfa Aesar) was added once. The tube was vortexed briefly (*ca.* 1 sec), inverted several times, and centrifuged for 1 min at 1,000 g. If there was foaming, an additional 10 μL of 1H,1H,2H,2H-perfluorooctanol was added. The aqueous layer was transferred to a fresh reaction tube and

the oil extracted twice more with 100 μL of the salmon sperm solution. The resulting 300 μL of recovered DNA solution was purified using Zymo Clean and Concentrate-5 kit (Zymo Research) according to the manufacturer's protocol. The DNA was eluted using 10 μL ultra-pure water after an extended incubation time of 10 min.

7.2 Biochemical procedures

7.2.1 General cloning procedures

Whenever commercial enzymes or kits were used, the manufacturer's instructions were followed unless otherwise indicated. Restriction enzymes were from the FastDigest series by Thermo Scientific unless otherwise indicated. Ligation reactions were performed using T4 DNA Ligase (Thermo Scientific). Plasmids were extracted from *E. coli* cells using the GeneJet Plasmid Miniprep Kit. Standard PCR was performed using Phusion High-Fidelity Polymerase (New England Biolabs) following the manufacturer's protocol. Colony PCR was performed using 2x DreamTaq Green MasterMix (ThermoScientific). Gel-electrophoresis was performed using 1% agarose gels containing 1x TAE buffer (40 mM Tris-base, 20 mM acetic acid, 1 mM EDTA, pH 8.5 at room temperature) and 1x SYBR Safe DNA gel stain (Invitrogen). Gels were typically run at 90 V for 30 min to 45 min. The DNA was visualised on a low-power UV-transilluminator (Vilber Lourmat), the required DNA band manually excised and the DNA extracted using the Zymoclean Gel DNA recovery kit (Zymo Research). If gel-electrophoresis was not required, DNA was purified using spin-columns (Zymo Clean and Concentrate-5 kit, Zymo Research). Electrocompetent cells were transformed using chilled 1 mm gap cuvettes and 1,800 V (Electroporator 2510, Eppendorf), followed by immediate addition of 1 mL of pre-warmed SOC medium (Lucigen) and recovery at 37 °C and 500 rpm for 60 min. The cells were plated onto LB agar plates containing the appropriate antibiotic(s). The concentrations of antibiotics were 100 $\mu\text{g } \mu\text{L}^{-1}$ for ampicillin (Amp), 30 $\mu\text{g } \mu\text{L}^{-1}$ kanamycin (Kan), 20 $\mu\text{g } \mu\text{L}^{-1}$ chloramphenicol (Cam) and 10 $\mu\text{g } \text{mL}^{-1}$ tetracycline (Tet), respectively.

7.2.2 Preparation of electro-competent cells

To each of two 2 L Erlenmeyer flasks containing 350 mL of pre-warmed SOB medium (20 g L^{-1} tryptone, 7 g L^{-1} yeast extract, 13 mM MgCl_2 , 13 mM MgSO_4 , 5 mM NaCl, 1 mM KCl) 3.5 mL of a densely grown pre-culture of *E. coli* strain BL21-Gold(DE3) (Agilent) in liquid LB+Tet. The cultures were incubated at 37 °C and 220 rpm until an OD_{600} of 0.6 was reached. Subsequent steps were conducted in a cold room (4 °C) using pre-chilled material. The cultures

were cooled in an ice-water bath for 15 min and centrifuged at 3,000 g and 4 °C for 10 min. The supernatant was discarded and the cell pellets resuspended in 350 mL of 1 mM HEPES (pH 7.0 at RT). This step was repeated twice with 15 min of centrifugation. Finally, the cell pellets were resuspended in 25 mL of 10% (w/v) glycerol each, combined and brought to 350 mL with 10% (w/v) glycerol. After a last centrifugation step the supernatant was discarded and the cell pellet slowly resuspended in 500 µL of 10% (w/v) glycerol. This cell suspension was divided into 50 µL aliquots, shock frozen in liquid nitrogen and stored at −80 °C. A test transformation with either ultra-pure resulted in a lawn on an LB+Tet agar plate and no colonies on an LB+Amp agar plate as expected. A test transformation with 10 ng pHAT5_HG3.17 showed an efficiency of 1×10^7 cfu/µg.

7.2.3 Construction of the metagenomic SCV library

The metagenomic libraries from soils, composts and vanilla pods were constructed by Dr Esther Gabor and Dr Annett Kirschner in Groningen. The cow rumen library was constructed by Dr Balint Kintsés and Dr Charlotte Miton in Cambridge. The SCV library combining all these libraries was prepared by Dr Pierre-Yves Colin in Cambridge. All libraries were stored at −20 °C.

The libraries constituting the SCV library were constructed from different soil samples [246], vanilla pods [261] and cow rumen [67], according to the method described by Gabor *et al.* [262]. Briefly, environmental DNA was sheared using a nebuliser, blunted using Klenow fragment and 3 to 5 kbp fragments were cloned into the EcoRV restriction site in plasmid pZero2 (Invitrogen). The DNA from cow rumen was partially digested with the restriction enzyme MluCI. DNA fragments in the size range of 3 kbp were isolated by gel electrophoresis and cloned into the EcoRI site of plasmid pZero2 (Invitrogen). Dr Pierre-Yves Colin received original library samples from the collaborators named above, pooled the ten libraries directly at ratios respecting the reported library sizes (Table B.1) creating the SCV library (oral communication Dr Pierre-Yves Colin). This pooled sample was directly used to transform cells for droplet screening in the work reported here, *i.e.* between the first construction to the screening in droplets there was one replication step in cells.

7.2.4 Construction of the epPCR Library of HG3.17

Random mutagenesis of HG3.17 was performed using the GeneMorph II kit (Agilent). Four reactions of 50 µL were prepared on ice, each containing 230 ng of plasmid pHAT5_HG3.17 (50 ng of target DNA), 1 µM of each T7 and T7t primers, 800 µM dNTPs and 1 µL of Mu-

tazyme II in 1x Mutazyme II reaction buffer. The PCR programme is detailed in Table 7.1. Two reactions were each subjected to 25 and 30 amplification cycles, respectively.

Table 7.1 PCR program for epPCR.

Segment	Cycles	Temperature (°C)	Duration (min)
1	1	95	2
2	25 or 30	95	0.5
		52	0.5
		72	1
3	1	72	10

After completion of the PCR programme, 0.5 µL of DpnI (FastDigest, Thermo Scientific) was added to each reaction to digest the template DNA, incubated for 60 min at 37 °C and inactivated for 5 min at 80 °C. The reaction mixtures were purified by gel-electrophoresis and the expected bands at 1 kbp extracted. The purified DNA was digested overnight (ca. 16 h) at 37 °C in 20 µL reactions using enzymes NcoI and XhoI (New England Biolabs), inactivated for 20 min at 80 °C and purified using spin-columns. The purified inserts were ligated at 3:1 molar excess into 25ng vector pHAT5 (previously digested with NcoI and XhoI) overnight (ca. 16 h) using T4 DNA ligase (Thermo Scientific) at 16 °C in 1x T4 DNA ligase buffer. After inactivation at 65 °C for 10 min, 2 µL of the mixture were transformed into 25 µL *E. coli* (E. cloni 10G Elite, Lucigen) and plated onto LB+Amp plates, number of transforming cells reported in Table 4.4.

To remove vector with no insert from the libraries, 1 µg of each was digested using EcoRI (FastDigest, Thermo Scientific) at 37 °C for 15 min, immediately inactivated at 80 °C for 5 min and purified using spin-columns. A mixture of the three libraries was prepared at ratios representing the original diversity resulting in a library with an estimated diversity of 1.5×10^5 . This epPCR library was used in the microfluidic droplet screening.

7.2.5 Construction of the deletion library of HG3.17

The HG3.17 gene was amplified from plasmid pET32-HG3.17 using the HG3.7-F and HG.317-B primers. The amplified gene was purified using gel-electrophoresis and digested using enzymes NdeI and XhoI (FastDigest, Thermo Scientific). The resulting insert was cloned into vector pID-tet (previously digested with NdeI and XhoI) to create plasmid pID-HG3.17. Vector pUC57-TransDel was digested using enzyme BglII using gel-electrophoresis and the expected band at 1.1 kbp extracted (TransDel transposon).

The transposon reaction was performed in 20 μL using 300ng of pID-HG3.17 and 50ng of the TransDel transposon using MuA transposase (ThermoScientific). The mixture was incubated for 1 h at 30 °C, inactivated immediately at 75 °C for 10 min, cooled on ice and purified using a spin-column. The purified DNA (1 μL) was transformed into *E. coli* (E. cloni 10G Elite, Lucigen) and plated onto one LB+Amp+Cam plate (5×10^4 transformants). The cells were scraped off using liquid LB and plasmids extracted. Of the resulting plasmids, 1 μg was digested using enzymes XhoI and NdeI and purified using gel-electrophoresis. The band corresponding to the target gene containing transposon (2 kbp) was extracted and ligated back into pID-tet vector overnight. The inactivated ligation was purified using a spin-column and 20 ng transformed into *E. coli* (E. cloni 10G Elite, Lucigen) and plated onto one LB+Amp+Cam agar plate (1.3×10^5 transformants). The cells were scraped from the plate and the plasmids extracted. Of the resulting pID-HG3.17xTransDel library, 1 μg was digested using MlyI and the expected DNA band purified from gel (2.7 kbp). The purified DNA was self-circularised using T4 DNA ligase overnight, purified using spin-columns and 20 ng, transformed into *E. coli* (E. cloni 10G Elite, Lucigen) and plated onto one LB+Amp agar plate (6.5×10^5 transformants). The cells were scraped off using liquid LB and the plasmids extracted resulting in the pID-HG3.17-Del library. Finally, the library was cloned into vector pHAT5 using the Eco72I and XhoI restriction sites giving the pHAT5-HG3.17-Del library used for droplet screening.

7.2.6 Construction of the insertion library of HG3.17

Up to restriction digestion by MlyI the insertion library was constructed following the protocol above but using the TransIns transposon from the pUC57-TransIns vector instead of TransDel. The transposon reaction with MuA yielded ca. 8.6×10^4 transformants. The second transformation step yielded ca. 1.3×10^5 transformants. Each 1 μg of the resulting plasmid library pID-H3.17xTransIns and the vector pUC57-Ins1 were digested using MlyI, purified by spin-column and eluted using 17 μL ultra-pure water. The eluate was used for a second restriction digestion using enzyme NotI. The DNA fragment excised from pUC57-Ins1 was ligated at 3:1 molar excess into 50 ng digested pID-H3.17xTransIns. The purified ligation product was transformed into *E. coli* (E. cloni 10G Elite, Lucigen) and plated onto one LB+Amp+Kan agar plate ($>1 \times 10^6$ transformants). The cells were scraped off the plates using liquid LB and the plasmids extracted.

Of the resulting library pID-HG3.17xIns1, 1 μg was digested using AclI for 40 min at 37 °C, immediately inactivated for 5 min at 65 °C and purified using a spin-column. Of the purified DNA, 1 μg was blunt-ended using 1U of the Large Klenow Fragment of DNA Polymerase 1 (New England Biolabs), by incubating for 15 min at 25 °C followed by immediate

inactivation by addition of 10 mM EDTA (final conc.) and incubation at 75 °C for 5 min. The product was gel purified and 50 ng self-circularised using 5U T4 DNA Ligase for 1 h at 22 °C and inactivated for 10 min at 75 °C. The DNA was purified using a spin-column, eluted using 7 µL of ultra-pure water and 3 µL transformed into *E. coli* (E. cloni 10G Elite, Lucigen). The cells were plate onto one LB+Amp agar plate (3 × 10⁶ transformants). The plasmids extracted resulting in the pID-HG3.17-Ins library. Finally, the library was cloned into vector pHAT5 using the Eco72I and XhoI restriction sites giving the pHAT5-HG3.17-Ins library used for droplet screening.

7.2.7 Expression of HG3.17 libraries for droplet screening

E. coli strain BL21-Gold(DE3) (Agilent) was transformed with 20 ng of the library and plated onto LB+Amp agar plates. The cells were grown overnight at 37 °C and scraped using 3x3 mL of liquid LB+Amp. The cell suspension from all plates were combined in the same reaction tube and gently mixed by inverting several times. Two liquid cultures were prepared by adding each 1 mL of cell suspension to 100 mL liquid LB+Amp. The cultures were incubated for 30 min at 20 °C and 220 rpm, followed by addition of IPTG to 400 µM (final conc.) and overnight incubation under these conditions. After protein expression, the cell suspensions were washed by pelleting the cells at 3,000 g and 4 °C, discarding the supernatant and resuspension in 50 mL reaction buffer (20 mM TrisHCl pH 7, 50 mM NaCl). This was done three times, on the last iteration the cells from the two separate liquid cultures were resuspended in 25 mL reaction buffer, combined and mixed by inverting the reaction tube several times. This cell suspension was used for droplet screening as described in Section 7.1.7.

7.2.8 Re-screening of HG3.17 variants in 96-well plates

Of the DNA recovered from the droplet sorting, 10 ng were transformed into electrocompetent *E. coli* (BL21-Gold(DE3), Agilent). Cells were plated onto LB+Amp agar aiming for 1,000 cfu / plate. The plates were incubated overnight at 37 °C and colonies picked using sterile pipettes to inoculate deep 96-well plates containing 500 µL liquid LB+Amp. The 96-well plates were incubated overnight at 37 °C and 750 rpm. From these plates, glycerol stocks were prepared by mixing 75 µL of cell culture and 25 µL of 50% w/v glycerol and followed by storage at –80 °C. Fresh plates were inoculated by adding 10 µL of the cell culture to 390 µL of liquid LB+Amp. The plates were incubated at 37 °C for 3 h, after which the temperature was reduced to 20 °C for 20 min followed by addition of 100 µL of 1.2 mM IPTG in LB+Amp to induce protein expression. Plates were incubated at 20 °C overnight. They were then centrifuged at 3,000 g for 1 h at 4 °C. The supernatant was discarded and the cell pellets frozen

at -20°C for 30 min. After thawing, 100 μL of freshly prepared lysis buffer was added (1x BugBuster (Merck), 15 μL of lysonase/50 mL (Merck) and 1 tablet Roche EDTA free protease inhibitor/50 mL). The cells were lysed at room temperature and 1,400 rpm for at least 20 min and stored at 4°C until used. The lysates were sequentially diluted by 10, 100 and 1,000 fold using an epMotion robot: 20 μL lysate was added to 180 μL reaction buffer (20 mM TrisHCl pH 7, 50 mM NaCl, filtered over 0.2 μm); the solution was mixed by pipetting up and down 10 times and 20 μL used for the next dilution step. Finally, 10 μL of the 1,000 fold dilution were added to 170 μL reaction buffer in a UV-transparent 96-well plate (Corning). The reaction was started by adding 20 μL of 20 mM 5-nitro-1,2-benzisoxazole in methanol and followed for 30 min at 380 nm with 15 sec resolution using a spectrophotometer (SpectraMax 190, Molecular Devices).

7.2.9 Protein expression and purification

The cells were grown in LB medium containing $100\text{ }\mu\text{g mL}^{-1}$ ampicillin at 37°C until an $\text{OD}_{600\text{nm}}$ of 0.5 to 0.7 was reached. Protein expression was induced by addition of 400 μM IPTG and the cells were grown at 20°C for 20 h. The cells were then centrifuged at 4,000 g for 10 min at 4°C . The cells were resuspended in 30 mL of 50 mM TrisHCl pH 8.0 and lysed using an Emulsiflex (Avestin). The extract was brought to 100 mM NaCl, 10 mM imidazole, and 10% v/v glycerol and centrifuged at 12,000 g for 40 min at 4°C to remove cell debris. The supernatant was loaded onto Ni-NTA resin (Qiagen) which was pre-equilibrated with purification buffer (50 mM TrisHCl pH 8.0, 100 mM NaCl, 10 mM imidazole). The proteins were eluted using a stepwise imidazole gradient (10, 50, 100, 200, 500 mM imidazole). Fractions containing the target protein were concentrated by centrifugation in an Amicon Ultra-15 concentrator with 30,000 MWCO (Merck-Millipore) at 3,000 g at 4°C . The samples were desalted over a Sephadex G-25 PD10 column (Amersham Biosciences) according to manufacturer's protocol and eluted using buffer with 50 mM TrisHCl pH 8.0 and 100 mM NaCl, 10% v/v glycerol. For the promiscuity test, the concentrated fractions were further purified by gel filtration on a HiLoad 16/60 Superdex 75 or 200 column (GE Healthcare) and fractions containing pure protein concentrated as above. Protein concentration was determined by measuring absorbance at 280 nm using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies). The same protocol was followed to purify HG3.17 and other variants, with 20 mM TrisHCl pH 7.0 and 50 mM NaCl instead of 50 mM TrisHCl pH 8.0 and 100 mM.

7.2.10 p-Nitrophenyl ester kinetics

A dilution series of the respective p-nitrophenyl carboxylate (all purchased from Sigma-Aldrich, except for p-nitrophenyl hexanoate, which was from TCI) in 180 μ L 50 mM TrisHCl pH 8.0 and 100 mM NaCl was prepared in the range of tens of micromolar to several millimolar. The reactions were performed in triplicate and started by addition of 20 μ L 10 \times purified enzyme in storage buffer or storage buffer only (blank) with the final concentration ranging from nanomolar to low micromolar. The reactions were followed at 405 nm for 15 min using a spectrophotometer (SpectraMax 190, Molecular Devices). Initial reaction rates were determined in the linear range and the blank subtracted.

7.2.11 Differential scanning fluorimetry

Enzymes were at 10 μ M in 50 mM Tris-HCl pH, 100 mM NaCl, and 2.5% glycerol with 2 \times or 5 \times SYPRO Orange Protein Gel Stain (from 5,000 \times concentrate in DMSO, Thermo Fisher). Samples of 20 μ L were measured in triplicates in reaction tubes which had been heated previously to reduce sticking of the dye to the plastic surface. Melting curves were recorded on a Rotor-Gene 6000 Real Time PCR Machine (Corbett). Fluorescence was excited at 470 nm and measured at 610 nm between 25 and 95 $^{\circ}$ C in steps of 0.5 $^{\circ}$ C.

7.2.12 Catalytic promiscuity test

All reactions were performed in 96 well plates, had a volume of 200 μ L, and were followed at 405 nm using a SpectraMax190 spectrophotometer (Molecular Devices). All enzymes were used at the highest concentration possible (N1ORF4: 1.6 μ M, N1ORF5: 13.7 μ M, N4: 13.8 μ M, N7: 28.8 μ M, N11: 0.3 μ M, N13: 0.3 μ M, N16: 14.7 μ M, N26: 35.7 μ M, RR11ORF1: 1.0 μ M, RR11ORF2: 55.1 μ M,). Each substrates was at 1 mM. The substrates were purchased from Sigma-Aldrich unless otherwise stated: 1: pNP-hexanoate, 2: pNP-dodecanoate, 3: oNP-decanoate, 4: β -lactamase substrate (CENTA[™], Calbiochem), 5: pNitroacetanilide, 6: pNP-phosphate, 7: pNP-phenylphosphonate, 8: pNP-sulfate, 9: pNP- β -D-glucopyranoside, 10: pNP- α -D-glucopyranoside, 11: pNP- β -D-galactopyranoside, 12: pNP- α -D-galactopyranoside, 13: pNP- β -D-xylopyranoside, 14: pNP- β -D-fucopyranoside, 15: 2-Chloro-4-phenyl- β -cellobioside (Megazyme), 16: pNP- β -xylobiose (Megazyme), 17: 5-nitrobenzisoxazole, 18: S-phenylthioacetate, 19: 2,3-Mercapto-1-propanol, 20: S-Methyl-thiobutanoate. In reactions 1-16 the buffer was 50 mM TrisHCl pH 8.0 and 100 mM NaCl. Reactions 18-20 were performed in the same buffer with 10 mM 5,5'-Dithiobis(2-nitrobenzoic acid) for the detection thiols [263]. Reaction 17 was performed in 20 mM TrisHCl pH 8.0 and 50 mM in TrisHCl pH 7.0.

7.2.13 Kemp elimination kinetics

The formation of 4-nitro-2-cyanophenol (Kemp reaction product of **2a**, see below) was monitored by UV–VIS spectroscopy (SpectraMax 190, Molecular Devices) at 380 nm at 25 °C for 30 min using UV-transparent 96-well plates (Corning). The reactions were set up by adding 20 μL of a 10x stock solution of purified enzyme in 20 mM TrisHCl pH 7, 50 mM to 160 μL of the same buffer and started by adding 20 μL of a 10x stock solution of the desired substrate concentration in MeOH (10% v/v final). The enzyme concentrations ranged between 5 and 10 nM and the substrate concentrations between 40 μM and 2 mM.

Buffer tests were performed using 40 mM buffer and 100 mM NaCl with 200 μM **2a**. The buffers were: sodium acetate (pH < 6), sodium phosphate (6 < pH < 8), sodium borate (pH 9) and sodium carbonate (pH 10). The substrate **2a** was dissolved at 10 mM in MeOH and diluted to 1 mM in 10 mM HCl. The reactions were started by the addition of 40 μL of the 1 mM substrate solution to each well. The total reaction volume was 200 μL .

7.3 Chemical procedures

7.3.1 Synthesis of 5-nitro-1,2-benzisoxazole **2a**

Concentrated sulphuric acid (5.4 mL, 89 mmol) was cooled in an ice/water bath to 0 °C and concentrated nitric acid (2.8 mL, 44 mmol) was added drop-wise. To this mixture, 1,2-benzisoxazole (3.0 mL, 30 mmol) was added drop-wise never allowing the reaction temperature to rise above 15 °C. The solution was stirred for 30 min and poured onto an ice/water mixture (1:1, 70 mL). The crude product was collected as a precipitate and recrystallised from anhydrous ethanol to yield 5-nitro-1,2-benzisoxazole (930 mg, 6 mmol, 20% yield) as colourless needles, δ_{H} (500 MHz; $(\text{CD}_3)_2\text{SO}$): 9.45 (s, 1 H), 8.88 (d, $J = 2.3$ Hz, 1 H), 8.50 (dd, $J = 9.3$ Hz, $J = 2.3$ Hz, 1 H), 8.02 (d, $J = 9.3$ Hz, 1H); δ_{C} (500 MHz; $(\text{CD}_3)_2\text{SO}$): 163.7 (arom. C-O), 148.7 (arom. CH=N), 144.4 (arom C-NO₂), 125.9, 122.1, 120.4, 110.7 (arom.).

7.3.2 Synthesis of 4-nitro-2-cyanophenol **2b**

5-Nitro-1,2-benzisoxazole was incubated under basic conditions (6.7 mM in 0.33 M NaOH) for three hours, followed by neutralisation with HCl yielding 5 mM 4-nitro-2-cyanophenol in 0.5 M NaCl. The apparent extinction coefficient was determined from four concentrations (50, 100, 150 and 200 μM) at each pH in the respective buffer with NaCl adjusted to 100 mM. A maximum was observed at 380 nm and the apparent extinction coefficient at pH 7 determined to be $\epsilon_{380\text{nm}} = 15,900 \text{ M}^{-1} \text{ cm}^{-1}$, in close agreement with literature [94].

7.3.3 Synthesis of 5-amino-1,2-benzisoxazole 5a

This compound was synthesised by Dr Josephin Holstein according to previously published procedures [191].

Yield 70%. δ_H (400 MHz; $CDCl_3$): 8.55 (s, 1H), 7.43 (d, $J = 8.8$ Hz, 1H), 6.98 (dd, $J = 2.3$ Hz, $J = 8.8$ Hz, 1H), 6.93 (dd, $J_1 = 0.5$ Hz, $J_2 = 1.8$ Hz, 1H). m/z for $C_7H_7N_2O^+$: 135.06 $[M-H]^+$; found: 135.2

7.3.4 Synthesis of 5-azido-1,2-benzisoxazole 6a

This compound was synthesised by Dr Josephin Holstein.

5-amino-1,2-benzisoxazole (100 mg, 0.75 mmol) was dissolved in 10 mL of 2 M HCl. The solution was cooled to 0 °C with an ice bath. An aqueous solution of $NaNO_2$ (65 mg, 0.78 mmol, in 1 mL water) was added slowly to the solution. The mixture was stirred for 30 min. Then, NaN_3 (95 mg, 1.5 mmol) in water (4.25 mL) was added to the mixture. The mixture was stirred for 3 h at room temperature. The resulting brownish solution was extracted with ethyl acetate (3×30 mL). The combined organic layers were washed with brine (50 mL). The solvent was evaporated and the resulting brown product dried under high vacuum yielding 60 mg (0.37 mmol, 50%) product. δ_H (400 MHz, $CD_3(SO)$): 8.69 (s, 1H), 7.63 (d, $J = 8.8$ Hz, 1H), 7.39 (d, $J = 2.2$ Hz, 1H), 7.26 (dd, $J = 2.2$ Hz, $J = 8.8$ Hz, 1H).

7.3.5 Small molecule characterisation

Dr Josephin Holstein lyophilised the reaction products and submitted them to the below services at the Department of Chemistry.

To characterise the reaction products of **6a**, it was converted using 1 μ M purified HG3.17 in 20 mM TrisHCl pH 7, 50 mM NaCl overnight in the dark. It was then either kept in the dark until measured or exposed to light of 365 nm for 10 min using a hand-held UV-lamp (ca. 100 μ Wcm⁻²). The enzyme was precipitated by addition formic acid (conc.) and the supernatant lyophilised (E-C MicroModulyo, ThermoFisher).

UPLC-MS Samples were dissolved in 50% aqueous acetonitrile and characterised using a Waters system consisting of an H-Class UPLC, a photodiode array for UV detection (190 to 800 nm) and a SQD2 single quadrupole mass spectrometer (100 to 500 m/z) using electrospray ionisation. Samples were run on a C18 UPLC column using a 5 to 95% acetonitrile gradient, 2 mM ammonium acetate, over 1 min under neutral, acidic (+ 0.1% formic acid) or

basic (+2 mM ammonium hydroxide) conditions with detection switching between positive and negative mode.

NMR The compound was dissolved in 600 μ L DMSO- d_6 (MagniSolve, Merck) and submitted to the NMR service at the Department of Chemistry where it was measured on a Bruker 400 MHz UltraShield Plus magnet equipped with a Quattro Nucleus Probe (QNP) cryoprobe.

7.4 Sequence Similarity Networks

The full-length fasta sequences were downloaded from Pfam (PFAM 32.0, <https://pfam.xfam.org/>, April 2019) for the largest families of each clan. To reduce the number of sequences (nodes) in the network, the individual families were clustered stepwise to 90%, 60% and 30% sequence identity using cd-hit and psi-cd-hit [264]. The most representative sequence for each cluster was used combined with the esterase hit sequences into a single database, which was used for an all-versus-all alignment (Protein-Protein BLAST 2.6.0+, [265]). Each line in the output file defined an edge (alignment) connecting two nodes (representative sequence) of the network. Duplicate edges, self-loops and edges above a certain e-value were removed using a custom python script. The simplified network was then imported into Cytoscape 3.7.1 to visualise the network [266].

Bibliography

- [1] A. Schmid, J. S. Dordick, B. Hauer, A. Kiener, M Wubbolt, and B. Witholt. Industrial Biocatalysis and Tomorrow. *Nature*, 409(6817):258–268, 2001.
- [2] Carlos Pedrós-Alió. The Rare Bacterial Biosphere. *Annual Review of Marine Science*, 4(1):449–466, 2012.
- [3] N H Horowitz. One-gene-one-enzyme: remembering biochemical genetics. *Protein Sci*, 4(5):1017–1019, 1995.
- [4] Olga Khersonsky and Dan S. Tawfik. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annual Review of Biochemistry*, 79(1):471–505, 2010.
- [5] Christopher K. Savile, Jacob M. Janey, Emily C. Mundorff, Jeffrey C. Moore, Sarena Tam, William R. Jarvis, Jeffrey C. Colbeck, Anke Krebber, Fred J. Fleitz, Jos Brands, Paul N. Devine, Gjalb W. Huisman, and Gregory J. Hughes. Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science*, 329(5989):305–309, 2010.
- [6] Patrick J. O’Brien and Daniel Herschlag. Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry and Biology*, 6(4):R91–R105, 1999.
- [7] Miguel D. Toscano, Kenneth J. Woycechowsky, and Donald Hilvert. Minimalist active-site redesign: Teaching old enzymes new tricks. *Angewandte Chemie - International Edition*, 46(18):3212–3236, 2007.
- [8] Nicholas J. Turner. Directed evolution drives the next generation of biocatalysts. *Nature Chemical Biology*, 5(8):567–573, 2009.
- [9] Cathleen Zeymer and Donald Hilvert. Directed Evolution of Protein Catalysts. *Annual Review of Biochemistry*, 87(1):131–157, 2018.
- [10] Hans Renata, Z. Jane Wang, and Frances H. Arnold. Expanding the enzyme universe: Accessing non-natural reactions by mechanism-guided directed evolution. *Angewandte Chemie - International Edition*, 54(11):3351–3367, 2015.
- [11] Daniela Röthlisberger, Olga Khersonsky, Andrew M. Wollacott, Lin Jiang, Jason DeChancie, Jamie Betker, Jasmine L. Gallaher, Eric A. Althoff, Alexandre Zanghellini, Orly Dym, Shira Albeck, Kendall N. Houk, Dan S. Tawfik, and David Baker. Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192):190–195, 2008.

- [12] Lin Jiang, Eric A. Althoff, Fernando R. Clemente, Lindsey Doyle, Daniela Röthlisberger, Alexandre Zanghellini, Jasmine L. Gallaher, Jamie L. Betker, Fujie Tanaka, Carlos F. Barbas, Donald Hilvert, Kendall N. Houk, Barry L. Stoddard, and David Baker. De novo computational design of retro-aldol enzymes. *Science*, 319(5868):1387–1391, 2008.
- [13] Justin B. Siegel, Alexandre Zanghellini, Helena M. Lovick, Gert Kiss, Abigail R. Lambert, Jennifer L. St.Clair, Jasmine L. Gallaher, Donald Hilvert, Michael H. Gelb, Barry L. Stoddard, Kendall N. Houk, Forrest E. Michael, and David Baker. Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science*, 329(5989):309–313, 2010.
- [14] Chi-Huey Wong and George M Whitesides. *Enzymes in Synthetic Organic Chemistry*, volume 12. Pergamon, 1994.
- [15] H. K. Privett, G. Kiss, T. M. Lee, R. Blomberg, R. A. Chica, L. M. Thomas, D. Hilvert, K. N. Houk, and S. L. Mayo. Iterative approach to computational enzyme design. *Proceedings of the National Academy of Sciences*, 109(10):3790–3795, 2012.
- [16] Rebecca Blomberg, Hajo Kries, Daniel M. Pinkas, Peer R E Mittl, Markus G. Grütter, Heidi K. Privett, Stephen L. Mayo, and Donald Hilvert. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature*, 503(7476):418–421, 2013.
- [17] Matthias Höhne, Sebastian Schätzle, Helge Jochens, Karen Robins, and Uwe T. Bornscheuer. Rational assignment of key motifs for function guides in silico enzyme identification. *Nature Chemical Biology*, 6(11):807–813, 2010.
- [18] Alex Bateman, Penny Coggill, and Robert D. Finn. DUFs: Families in search of function. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 66(10):1148–1152, 2010.
- [19] Robert D. Finn, Jody Clements, William Arndt, Benjamin L. Miller, Travis J. Wheeler, Fabian Schreiber, Alex Bateman, and Sean R. Eddy. HMMER web server: 2015 Update. *Nucleic Acids Research*, 43(W1):W30–W38, 2015.
- [20] William R Pearson. An introduction to sequence similarity (“homology”) searching. *Current Protocols in Bioinformatics*, Chapter 3(SUPPL.42):Unit3.1, 2013.
- [21] Yi Ling Du, Hai Yan He, Melanie A. Higgins, and Katherine S. Ryan. A heme-dependent enzyme forms the nitrogen-nitrogen bond in piperazate. *Nature Chemical Biology*, 13(8):836–838, 2017.
- [22] Anonymous. Hints of hidden chemistry. *Nature Chemical Biology*, 13(8):815, 2017.
- [23] J. Staley. Measurement of In Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. *Annual Review of Microbiology*, 39(1):321–346, 1985.
- [24] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

- [25] Norman R. Pace, David A. Stahl, David J. Lane, and Gary J. Olsen. The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. pages 1–55. Springer, Boston, MA, 1986.
- [26] T. M. Schmidt, E. F. DeLong, and N. R. Pace. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology*, 173(14):4371–4378, 1991.
- [27] J. Craig Venter, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug Rusch, Jonathan A. Eisen, Dongying Wu, Ian Paulsen, Karen E. Nelson, William Nelson, Derrick E. Fouts, Samuel Levy, Anthony H. Knap, Michael W. Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu Hui Rogers, and Hamilton O. Smith. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004.
- [28] Supratim Mukherjee, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Olena Verezemskaya, Michelle Isbandi, Alex D. Thomas, Rida Ali, Kaushal Sharma, Nikos C. Kyrpides, and T. B.K. Reddy. Genomes OnLine Database (GOLD) v.6: Data updates and feature enhancements. *Nucleic Acids Research*, 45(D1):D446–D456, 2017.
- [29] O. Beja, L. Aravind, E. V. Koonin, M. T. Suzuki, A. Hadd, L. P. Nguyen, S. B. Jovanovich, C. M. Gates, R. A. Feldman, J. L. Spudich, E. N. Spudich, and E. F. DeLong. Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science*, 289(5486):1902–1906, 2000.
- [30] Alexander H. Treusch, Sven Leininger, Arnulf Kietzin, Stephan C. Schuster, Hans Peter Klenk, and Christa Schleper. Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. *Environmental Microbiology*, 7(12):1985–1995, 2005.
- [31] Manuel Ferrer, Mónica Martínez-Martínez, Rafael Bargiela, Wolfgang R. Streit, Olga V. Golyshina, and Peter N. Golyshin. Estimating the success of enzyme bioprospecting through metagenomics: Current status and future trends. *Microbial Biotechnology*, 9(1):22–34, 2016.
- [32] Jo Handelsman, Michelle R. Rondon, Sean F. Brady, Jon Clardy, and Robert M. Goodman. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry and Biology*, 5(10):R245–R249, 1998.
- [33] Carola Simon and Rolf Daniel. Achievements and new knowledge unraveled by metagenomic approaches. *Applied Microbiology and Biotechnology*, 85(2):265–276, 2009.
- [34] F. G. Healy, R. M. Ray, H. C. Aldrich, A. C. Wilkie, L. O. Ingram, and K. T. Shanmugam. Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. *Applied Microbiology and Biotechnology*, 43(4):667–674, 1995.
- [35] Rolf Daniel. The metagenomics of soil. *Nature Reviews Microbiology*, 3(6):470–478, 2005.

- [36] Esther M. Gabor, Wynand B.L. Alkema, and Dick B. Janssen. Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environmental Microbiology*, 6(9):879–886, 2004.
- [37] Kathy N. Lam, Jiujuun Cheng, Katja Engel, Josh D. Neufeld, and Trevor C. Charles. Current and future resources for functional metagenomics. *Frontiers in Microbiology*, 6(OCT):1196, 2015.
- [38] Christian Leggewie, Helge Henning, Christel Schmeisser, Wolfgang R. Streit, and Karl Erich Jaeger. A novel transposon for functional expression of DNA libraries. *Journal of Biotechnology*, 123(3):281–287, 2006.
- [39] Yu Jung Kim, Haseong Kim, Seo Hyeon Kim, Eugene Rha, Su Lim Choi, Soo Jin Yeom, Hak Sung Kim, and Seung Goo Lee. Improved metagenome screening efficiency by random insertion of T7 promoters. *Journal of Biotechnology*, 230:47–53, 2016.
- [40] Katrin Lämmle, Hubert Zipper, Michael Breuer, Bernhard Hauer, Christiane Buta, Herwig Brunner, and Steffen Rupp. Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. *Journal of Biotechnology*, 127(4):575–592, 2007.
- [41] Stefan M. Gaida, Nicholas R. Sandoval, Sergios A. Nicolaou, Yili Chen, Keerthi P. Venkataramanan, and Eleftherios T. Papoutsakis. Expression of heterologous sigma factors enables functional screening of metagenomic and heterologous genomic libraries. *Nature Communications*, 6:7045, 2015.
- [42] Esther Gabor, Klaus Liebeton, Frank Niehaus, Juergen Eck, and Patrick Lorenz. Updating the metagenomics toolbox. *Biotechnology Journal*, 2(2):201–206, 2007.
- [43] Nadine Katzke, Andreas Knapp, Anita Loeschcke, Thomas Drepper, and Karl Erich Jaeger. Novel tools for the functional expression of metagenomic DNA. In *Methods in Molecular Biology*, volume 1539, pages 159–196. Humana Press, Totowa, NJ, 2017.
- [44] Benedikt Leis, Angel Angelov, Markus Mientus, Haijuan Li, Vu T.T. Pham, Benjamin Lauinger, Patrick Bongen, Jörg Pietruszka, Luís G. Gonçalves, Helena Santos, and Wolfgang Liebl. Identification of novel esterase-active enzymes from hot environments by use of the host bacterium *Thermus thermophilus*. *Frontiers in Microbiology*, 6(APR):275, 2015.
- [45] Vinayak Agarwal, Jessica M. Blanton, Sheila Podell, Arnaud Taton, Michelle A. Schorn, Julia Busch, Zhenjian Lin, Eric W. Schmidt, Paul R. Jensen, Valerie J. Paul, Jason S. Biggs, James W. Golden, Eric E. Allen, and Bradley S. Moore. Metagenomic discovery of polybrominated diphenyl ether biosynthesis by marine sponges. *Nature Chemical Biology*, 13(5):537–543, 2017.
- [46] Philip Mair, Fabrice Gielen, and Florian Hollfelder. Exploring sequence space in search of functional enzymes using microfluidic droplets. *Current Opinion in Chemical Biology*, 37:137–144, 2017.

- [47] Richard Obexer, Moritz Pott, Cathleen Zeymer, Andrew D. Griffiths, and Donald Hilvert. Efficient laboratory evolution of computationally designed enzymes with low starting activities using fluorescence-activated droplet sorting. *Protein Engineering, Design and Selection*, 29(9):355–366, 2016.
- [48] Lars Giger, Sami Caner, Richard Obexer, Peter Kast, David Baker, Nenad Ban, and Donald Hilvert. Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nature Chemical Biology*, 9(8):494–498, 2013.
- [49] Yolanda Schaerli and Florian Hollfelder. The potential of microfluidic water-in-oil droplets in experimental biology. *Molecular BioSystems*, 5(12):1392–1404, 2009.
- [50] Liisa D. van Vliet, Pierre Yves Colin, and Florian Hollfelder. Bioinspired genotype–phenotype linkages: Mimicking cellular compartmentalization for the engineering of functional proteins. *Interface Focus*, 5(4), 2015.
- [51] David J. Collins, Adrian Neild, Andrew DeMello, Ai Qun Liu, and Ye Ai. The Poisson distribution and beyond: Methods for microfluidic droplet production and single cell encapsulation. *Lab on a Chip*, 15(17):3439–3459, 2015.
- [52] Dan S. Tawfik and Andrew D. Griffiths. Man-made cell-like compartments for molecular evolution. *Nature Biotechnology*, 16(7):652–656, 1998.
- [53] Pierre Yves Colin, Anastasia Zinchenko, and Florian Hollfelder. Enzyme engineering in biomimetic compartments. *Current Opinion in Structural Biology*, 33:42–51, 2015.
- [54] Samuel C. Kim, Gayatri Premasekharan, Iain C. Clark, Hawi B. Gameda, Pamela L. Paris, and Adam R. Abate. Measurement of copy number variation in single cancer cells using rapid-emulsification digital droplet MDA. *Microsystems & Nanoengineering*, 3:17018, 2017.
- [55] Younan Xia and George M. Whitesides. Soft Lithography. *Angewandte Chemie International Edition*, 37(5):550–575, 1998.
- [56] A. Del Campo and C Greiner. SU-8: A photoresist for high-aspect-ratio and 3D sub-micron lithography. *Journal of Micromechanics and Microengineering*, 17(6):R81–R95, 2007.
- [57] Todd Thorsen, Richard W. Roberts, Frances H. Arnold, and Stephen R. Quake. Dynamic Pattern Formation in a Vesicle-Generating Microfluidic Device. *Physical Review Letters*, 86(18):4163–4166, 2001.
- [58] Balint Kintses, Liisa D. van Vliet, Sean R.A. Devenish, and Florian Hollfelder. Microfluidic droplets: New integrated workflows for biological experiments. *Current Opinion in Chemical Biology*, 14(5):548–555, 2010.
- [59] A. R. Abate, T. Hung, P. Mary, J. J. Agresti, and D. A. Weitz. High-throughput injection with microfluidics using picoinjectors. *Proceedings of the National Academy of Sciences*, 107(45):19163–19166, 2010.

- [60] Jean Christophe Baret, Oliver J. Miller, Valerie Taly, Michaël Ryckelynck, Abdelslam El-Harrak, Lucas Frenz, Christian Rick, Michael L. Samuels, J. Brian Hutchison, Jeremy J. Agresti, Darren R. Link, David A. Weitz, and Andrew D. Griffiths. Fluorescence-activated droplet sorting (FADS): Efficient microfluidic cell sorting based on enzymatic activity. *Lab on a Chip*, 9(13):1850–1858, 2009.
- [61] Fabrice Gielen, Raphaëlle Hours, Stephane Emond, Martin Fischlechner, Ursula Schell, and Florian Hollfelder. Ultrahigh-throughput-directed enzyme evolution by absorbance-activated droplet sorting (AADS). *Proceedings of the National Academy of Sciences*, 113(47):E7383–E7389, 2016.
- [62] Anastasia Zinchenko, Sean R A Devenish, Balint Kintses, Pierre Yves Colin, Martin Fischlechner, and Florian Hollfelder. One in a million: Flow cytometric sorting of single cell-lysate assays in monodisperse picolitre double emulsion droplets for directed evolution. *Analytical Chemistry*, 86(5):2526–2533, 2014.
- [63] Martin Fischlechner, Yolanda Schaerli, Mark F. Mohamed, Santosh Patil, Chris Abell, and Florian Hollfelder. Evolution of enzyme catalysts caged in biomimetic gel-shell beads. *Nature Chemistry*, 6(9):791–796, 2014.
- [64] J. J. Agresti, E. Antipov, A. R. Abate, K. Ahn, A. C. Rowat, J.-C. Baret, M. Marquez, A. M. Klibanov, A. D. Griffiths, and D. A. Weitz. Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proceedings of the National Academy of Sciences*, 107(9):4004–4009, 2010.
- [65] Rui Wang, Weihong Zheng, Haiqiang Yu, Haiteng Deng, and Minkui Luo. Labeling Substrates of Protein Arginine Methyltransferase with. *Journal of the American Chemical Society*, 133(20):7648–7651, 2011.
- [66] Majdi Najah, Raphaël Calbrix, I. Putu Mahendra-Wijaya, Thomas Beneyton, Andrew D. Griffiths, and Antoine Drevelle. Droplet-based microfluidics platform for ultra-high-throughput bioprospecting of cellulolytic microorganisms. *Chemistry and Biology*, 21(12):1722–1732, 2014.
- [67] Pierre Yves Colin, Balint Kintses, Fabrice Gielen, Charlotte M. Miton, Gerhard Fischer, Mark F. Mohamed, Marko Hyvönen, Diego P. Morgavi, Dick B. Janssen, and Florian Hollfelder. Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nature Communications*, 6(1):10008, 2015.
- [68] Staffan L. Sjöström, Yunpeng Bai, Mingtao Huang, Zihe Liu, Jens Nielsen, Haakan N. Joensson, and Helene Andersson Svahn. High-throughput screening for industrial enzyme production hosts by droplet microfluidics. *Lab on a Chip*, 14(4):806–813, 2014.
- [69] Balint Kintses, Christopher Hein, Mark F. Mohamed, Martin Fischlechner, Fabienne Courtois, Céline Lainé, and Florian Hollfelder. Picoliter cell lysate assays in microfluidic droplet compartments for directed enzyme evolution. *Chemistry and Biology*, 19(8):1001–1009, 2012.
- [70] Anastasia Zinchenko. *Development of a user-friendly high-throughput screening system in microdroplets for the selection of efficient biocatalysts*. Doctoral thesis., University of Cambridge, 2015.

- [71] Philipp Gruner, Birte Riechers, Laura Andreina Chacòn Orellana, Quentin Brosseau, Florine Maes, Thomas Beneyton, Deniz Pekin, and Jean Christophe Baret. Stabilisers for water-in-fluorinated-oil dispersions: Key properties for microfluidic applications. *Current Opinion in Colloid and Interface Science*, 20(3):183–191, 2015.
- [72] Philipp Gruner, Birte Riechers, Benoît Semin, Jiseok Lim, Abigail Johnston, Kathleen Short, and Jean-Christophe Baret. Controlling molecular transport in minimal emulsions. *Nature Communications*, 7:9, 2016.
- [73] Fabienne Courtois, Luis F. Olguin, Graeme Whyte, Ashleigh B. Theberge, Wilhelm T S Huck, Florian Hollfelder, and Chris Abell. Controlling the retention of small molecules in emulsion microdroplets for use in cell-based assays. *Analytical Chemistry*, 81(8):3008–3016, 2009.
- [74] Majdi Najah, Estelle Mayot, I. Putu Mahendra-Wijaya, Andrew D. Griffiths, Sylvain Ladame, and Antoine Drevelle. New glycosidase substrates for droplet-based microfluidic screening. *Analytical Chemistry*, 85(20):9807–9814, 2013.
- [75] Johan Fenneteau, Dany Chauvin, Andrew D. Griffiths, Clément Nizak, and Janine Cossy. Synthesis of new hydrophilic rhodamine based enzymatic substrates compatible with droplet-based microfluidic assays. *Chemical Communications*, 53(39):5437–5440, 2017.
- [76] Fuqiang Ma, Michael Fischer, Yunbin Han, Stephen G. Withers, Yan Feng, and Guang Yu Yang. Substrate Engineering Enabling Fluorescence Droplet Entrapment for IVC-FACS-Based Ultrahigh-Throughput Screening. *Analytical Chemistry*, 88(17):8587–8595, 2016.
- [77] Raluca Ostafe, Radivoje Prodanovic, W. Lloyd Ung, David A. Weitz, and Rainer Fischer. A high-throughput cellulase screening system based on droplet microfluidics. *Biomicrofluidics*, 8(4):041102, 2014.
- [78] Thomas Beneyton, I. Putu Mahendra Wijaya, Prexilia Postros, Majdi Najah, Pascal Leblond, Angelique Couvent, Estelle Mayot, Andrew D. Griffiths, and Antoine Drevelle. High-throughput screening of filamentous fungi using nanoliter-range droplet-based microfluidics. *Scientific Reports*, 6(January):27223, 2016.
- [79] Richard Obexer, Alexei Godina, Xavier Garrabou, Peer R.E. Mittl, David Baker, Andrew D. Griffiths, and Donald Hilvert. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nature Chemistry*, 9(1):50–56, 2017.
- [80] Andrew C. Larsen, Matthew R. Dunn, Andrew Hatch, Sujay P. Sau, Cody Youngbull, and John C. Chaput. A general strategy for expanding polymerase function by droplet microfluidics. *Nature Communications*, 7:11235, 2016.
- [81] Michael Ryckelynck, Stéphanie Baudrey, Christian Rick, Annick Marin, Faith Coldren, Eric Westhof, and Andrew D. Griffiths. Using droplet-based microfluidics to improve the catalytic properties of RNA under multiple-turnover conditions. *Rna*, 21(3):458–469, 2015.

- [82] Ee Xien Ng, Miles A. Miller, Tengyang Jing, and Chia Hung Chen. Single cell multiplexed assay for proteolytic activity using droplet microfluidics. *Biosensors and Bioelectronics*, 81:408–414, 2016.
- [83] Alexander K. Price, Andrew B. MacConnell, and Brian M. Paegel. HvSABR: Photochemical Dose-Response Bead Screening in Droplets. *Analytical Chemistry*, 88(5):2904–2911, 2016.
- [84] Pierre-yves Colin. - *From Metagenomes to Directed Evolution - Microfluidic droplets identify novel promiscuous enzymes in environmental gene libraries*. Doctoral thesis, University of Cambridge, 2015.
- [85] Rainer Spang and Martin Vingron. Limits of homology detection by pairwise sequence comparison. *Bioinformatics*, 17(4):338–342, 2001.
- [86] Kristoffer Illergård, David H. Ardell, and Arne Elofsson. Structure is three to ten times more conserved than sequence - A study of structural response in protein cores. *Proteins: Structure, Function and Bioinformatics*, 77(3):499–508, 2009.
- [87] Elisa Lanfranchi, Tea Pavkov-Keller, Eva Maria Koehler, Matthias Diepold, Kerstin Steiner, Barbara Darnhofer, Jürgen Hartler, Tom Van Den Bergh, Henk Jan Joosten, Mandana Gruber-Khadjawi, Gerhard G Thallinger, Ruth Birner-Gruenberger, Karl Gruber, Margit Winkler, and Anton Glieder. Enzyme discovery beyond homology: A unique hydroxynitrile lyase in the Bet v1 superfamily. *Scientific Reports*, 7:46738, 2017.
- [88] Jennifer L Seffernick, Mervyn L de Souza, Michael J Sadowsky, and Lawrence P Wackett. Melamine Deaminase and Atrazine Chlorohydrolase: 98 Percent Identical but Functionally Different. *Journal of Bacteriology*, 183(8):2405 LP – 2410, 2001.
- [89] R A Jensen. Enzyme Recruitment in Evolution of New Function. *Annual Review of Microbiology*, 30(1):409–425, 1976.
- [90] Matthew A. Oberhardt, Raphy Zarecki, Leah Reshef, Fangfang Xia, Miquel Duran-Frigola, Rachel Schreiber, Christopher S. Henry, Nir Ben-Tal, Daniel J. Dwyer, Uri Gophna, and Eytan Ruppin. Systems-Wide Prediction of Enzyme Promiscuity Reveals a New Underground Alternative Route for Pyridoxal 5'-Phosphate Production in *E. coli*. *PLoS Computational Biology*, 12(1):e1004705, 2016.
- [91] Gabriela I. Guzman, Troy E. Sandberg, Ryan A. LaCroix, Akos Nyerges, Henrietta Papp, Markus de Raad, Zachary A. King, Trent R. Northen, Richard A. Notebaart, Csaba Pal, Bernhard O. Palsson, Balazs Papp, and Adam M. Feist. Enzyme promiscuity shapes evolutionary innovation and optimization. *bioRxiv*, page 310946, 2018.
- [92] Shelley D. Copley. Evolution of efficient pathways for degradation of anthropogenic chemicals. *Nature Chemical Biology*, 5(8):559–566, 2009.
- [93] Carola Simon and Rolf Daniel. Metagenomic analyses: Past and future trends. *Applied and Environmental Microbiology*, 77(4):1153–1161, 2011.

- [94] Martha L. Casey, D. S. Kemp, Kenneth G. Paul, and Daniel D. Cox. The Physical Organic Chemistry of Benzisoxazoles. I. The Mechanism of the Base-Catalyzed Decomposition of Benzisoxazoles. *Journal of Organic Chemistry*, 38(13):2294–2301, 1973.
- [95] D. S. Kemp and Martha L. Casey. Physical Organic Chemistry of Benzisoxazoles. II. Linearity of the Brønsted Free Energy Relationship for the Base-Catalyzed Decomposition of Benzisoxazoles. *Journal of the American Chemical Society*, 95(20):6670–6680, 1973.
- [96] William J. Dower, Jeff F. Miller, and Charles W. Ragsdale. High efficiency transformation of E.coli by high voltage electroporation. *Nucleic Acids Research*, 16(13):6127–6145, 1988.
- [97] Shelley L. Anna, Nathalie Bontoux, and Howard A. Stone. Formation of dispersions using “flow focusing” in microchannels. *Applied Physics Letters*, 82(3):364–366, 2003.
- [98] Levent Yobas, Stefan Martens, Wee Liat Ong, and Nagarajan Ranganathan. High-performance flow-focusing geometry for spontaneous generation of monodispersed droplets. *Lab on a Chip*, 6(8):1073–1079, 2006.
- [99] Heng Dong Xi, Hao Zheng, Wei Guo, Alfonso M. Gañán-Calvo, Ye Ai, Chia Wen Tsao, Jun Zhou, Weihua Li, Yanyi Huang, Nam Trung Nguyen, and Say Hwa Tan. Active droplet sorting in microfluidics: a review. *Lab on a Chip*, 17(5):751–771, 2017.
- [100] Thomas Franke, Adam R. Abate, David A. Weitz, and Achim Wixforth. Surface acoustic wave (SAW) directed droplet flow in microfluidics for PDMS devices. *Lab on a Chip*, 9(18):2625–2627, 2009.
- [101] Lothar Schmid, David A. Weitz, and Thomas Franke. Sorting drops and cells with acoustics: Acoustic microfluidic fluorescence-activated cell sorter. *Lab on a Chip*, 14(19):3710–3718, 2014.
- [102] Adam Sciambi and Adam R. Abate. Generating electric fields in PDMS microfluidic devices with salt water electrodes. *Lab on a Chip*, 14(15):2605–2609, 2014.
- [103] Keunho Ahn, Charles Kerbage, Tom P. Hunt, R. M. Westervelt, Darren R. Link, and D. A. Weitz. Dielectrophoretic manipulation of drops for high-speed microfluidic sorting devices. *Applied Physics Letters*, 88(2):1–3, 2006.
- [104] Peter R.C. Gascoyne and Jody Vykoukal. Particle separation by dielectrophoresis, 2002.
- [105] Maria Tenje, Anna Fornell, Mathias Ohlin, and Johan Nilsson. Particle Manipulation Methods in Droplet Microfluidics. *Analytical Chemistry*, 90(3):1434–1443, 2018.
- [106] Ronald Pethig. How does Dielectrophoresis Differ from Electrophoresis? In *Dielectrophoresis*, pages 31–48. John Wiley & Sons, Ltd, Chichester, UK, 2017.
- [107] Electronics Materials Solutions Division. Heat transfer applications using 3M™ Novec™ Engineered Fluids. Technical report, 3M, 2016.

- [108] Adam Sciambi and Adam R. Abate. Accurate microfluidic sorting of droplets at 30 kHz. *Lab on a Chip*, 15(1):47–51, 2015.
- [109] Carl A. Heller, Ronald A. Henry, Bruce A. McLaughlin, and Dan E. Bliss. Fluorescence Spectra and Quantum Yields: Quinine, Uranine, 9,10-Diphenylanthracene, and 9,10-Bis(phenylethynyl)anthracenes. *Journal of Chemical and Engineering Data*, 19(3):214–219, 1974.
- [110] Yoshihiro Shimizu, Takashi Kanamori, and Takuya Ueda. Protein synthesis by pure translation systems. *Methods*, 36(3):299–304, 2005.
- [111] Slawomir Jakiela, Sylwia Makulska, Piotr M. Korczyk, and Piotr Garstecki. Speed of flow of individual droplets in microfluidic channels as a function of the capillary number, volume of droplets and contrast of viscosities. *Lab on a Chip*, 11(21):3603–3608, 2011.
- [112] Joachim Jose. Autodisplay: Efficient bacterial surface display of recombinant proteins. *Applied Microbiology and Biotechnology*, 69(6):607–614, 2006.
- [113] Romas J. Kazlauskas, Alexandra N.E. Weissfloch, Aviva T. Rappaport, and Louis A. Cuccia. A Rule To Predict Which Enantiomer of a Secondary Alcohol Reacts Faster in Reactions Catalyzed by Cholesterol Esterase, Lipase from *Pseudomonas cepacia*, and Lipase from *Candida rugosa*. *Journal of Organic Chemistry*, 56(8):2656–2665, 1991.
- [114] Mónica Martínez-Martínez, Cristina Coscolín, Gerard Santiago, Jennifer Chow, Peter J. Stogios, Rafael Bargiela, Christoph Gertler, José Navarro-Fernández, Alexander Bollinger, Stephan Thies, Celia Méndez-García, Ana Popovic, Greg Brown, Tatyana N. Chernikova, Antonio García-Moyano, Gro E.K. Bjerga, Pablo Pérez-García, Tran Hai, Mercedes V. Del Pozo, Runar Stokke, Ida H. Steen, Hong Cui, Xiaohui Xu, Boguslaw P. Nocek, María Alcaide, Marco Distaso, Victoria Mesa, Ana I. Peláez, Jesús Sánchez, Patrick C.F. Buchholz, Jürgen Pleiss, Antonio Fernández-Guerra, Frank O. Glöckner, Olga V. Golyshina, Michail M. Yakimov, Alexei Savchenko, Karl Erich Jaeger, Alexander F. Yakunin, Wolfgang R. Streit, Peter N. Golyshin, Víctor Guallar, and Manuel Ferrer. Determinants and Prediction of Esterase Substrate Promiscuity Patterns. *ACS Chemical Biology*, 13(1):225–234, 2018.
- [115] Amrita Srivastava and Mayuri Srivastava. Global Enzymes Market. Opportunity Analysis and Industry Forecast, 2018-2024. Technical report, Allied Market Research, 2018.
- [116] Uwe T. Bornscheuer. Enzymes in Lipid Modification. *Annual Review of Food Science and Technology*, 9(1):85–103, 2018.
- [117] Uwe T. Bornscheuer. Microbial carboxyl esterases: Classification, properties and application in biocatalysis. *FEMS Microbiology Reviews*, 26(1):73–81, 2002.
- [118] José A.M. Prates, Nicolas Tarbouriech, Simon J. Charnock, Carlos M.G.A. Fontes, Luís M.A. Ferreira, and Gideon J. Davies. The structure of the feruloyl esterase module of xylanase 10B from *Clostridium thermocellum* provides insights into substrate recognition. *Structure*, 9(12):1183–1190, 2001.

- [119] Alissa Rauwerdink and Romas J. Kazlauskas. How the Same Core Catalytic Machinery Catalyzes 17 Different Reactions: The Serine-Histidine-Aspartate Catalytic Triad of α/β -Hydrolase Fold Enzymes. *ACS Catalysis*, 5(10):6153–6176, 2015.
- [120] Robert D. Finn, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A. Salazar, John Tate, and Alex Bateman. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285, 2016.
- [121] C. Elend, C. Schmeisser, C. Leggewie, P. Babiak, J. D. Carballeira, H. L. Steele, J. L. Reymond, K. E. Jaeger, and W. R. Streit. Isolation and biochemical characterization of two novel metagenome-derived esterases. *Applied and Environmental Microbiology*, 72(5):3637–3645, 2006.
- [122] Mariana Rangel Pereira, Gustavo Fernando Mercaldi, Thaís Carvalho Maester, Andrea Balan, and Eliana Gertrudes De Macedo Lemos. Est16, a new esterase isolated from a metagenomic library of a microbial consortium specializing in diesel oil degradation. *PLoS ONE*, 10(7):e0133723, 2015.
- [123] Florence Privé, C. Jamie Newbold, Naheed N. Kaderbhai, Susan G. Girdwood, Olga V. Golyshina, Peter N. Golyshin, Nigel D. Scollan, and Sharon A. Huws. Isolation and characterization of novel lipases/esterases from a bovine rumen metagenome. *Applied Microbiology and Biotechnology*, 99(13):5475–5485, 2015.
- [124] María Cecilia Rodríguez, Inés Loaces, Vanesa Amarelle, Daniella Senatore, Andrés Iriarte, Elena Fabiano, and Francisco Noya. Est10: A novel alkaline esterase isolated from bovine rumen belonging to the new family XV of lipolytic enzymes. *PLoS ONE*, 10(5):e0126651, 2015.
- [125] Stephan Thies, Sonja Christina Rausch, Filip Kovacic, Alexandra Schmidt-Thaler, Susanne Wilhelm, Frank Rosenau, Rolf Daniel, Wolfgang Streit, Jörg Pietruszka, and Karl Erich Jaeger. Metagenomic discovery of novel enzymes and biosurfactants in a slaughterhouse biofilm microbial community. *Scientific Reports*, 6(1):27035, 2016.
- [126] Mariana Rangel Pereira, Thaís Carvalho Maester, Gustavo Fernando Mercaldi, Eliana Gertrudes de Macedo Lemos, Marko Hyvönen, and Andrea Balan. From a metagenomic source to a high-resolution structure of a novel alkaline esterase. *Applied Microbiology and Biotechnology*, 101(12):4935–4949, 2017.
- [127] Ana Popovic, Tran Hai, Anatoly Tchigvintsev, Mahbod Hajighasemi, Boguslaw Nock, Anna N. Khusnutdinova, Greg Brown, Julia Glinos, Robert Flick, Tatiana Skarina, Tatyana N. Chernikova, Veronica Yim, Thomas Bröls, Denis Le Paslier, Michail M. Yakimov, Andrzej Joachimiak, Manuel Ferrer, Olga V. Golyshina, Alexei Savchenko, Peter N. Golyshin, and Alexander F. Yakunin. Activity screening of environmental metagenomic libraries reveals novel carboxylesterase families. *Scientific Reports*, 7:44103, 2017.
- [128] Erik Borchert, Joseph Selvin, Seghal G. Kiran, Stephen A. Jackson, Fergal O’Gara, and Alan D. W. Dobson. A Novel Cold Active Esterase from a Deep Sea Sponge *Stelletta normani* Metagenomic Library. *Frontiers in Marine Science*, 4:287, 2017.

- [129] Yuliya V. Samoylova, Ksenia N. Sorokina, Margarita V. Romanenko, and Valentin N. Parmon. Cloning, expression and characterization of the esterase estUT1 from *Ureibacillus thermosphaericus* which belongs to a new lipase family XVIII. *Extremophiles*, 22(2):271–285, 2018.
- [130] Sally Bayer, Anja Kunert, Meike Ballschmiter, and Thomas Greiner-Stoeffe. Indication for a new lipolytic enzyme family: Isolation and characterization of two esterases from a metagenomic library. *Journal of Molecular Microbiology and Biotechnology*, 18(3):181–187, 2010.
- [131] Jay M. Short, Joseph M. Fernandez, Joseph A. Sorge, and William D. Huse. λ ZAP: A bacteriophage λ expression vector with in vivo excision properties. *Nucleic Acids Research*, 16(15):7583–7600, 1988.
- [132] Manuel Ferrer, Rafael Bargiela, Mónica Martínez-Martínez, Jaume Mir, Rainhard Koch, Olga V. Golyshina, and Peter N. Golyshin. Biodiversity for biocatalysis: A review of the α/β -hydrolase fold superfamily of esterases-lipases discovered in metagenomes. *Biocatalysis and Biotransformation*, 33(5-6):235–249, 2015.
- [133] Nicolas Lenfant, Thierry Hotelier, Eric Velluet, Yves Bourne, Pascale Marchot, and Arnaud Chatonnet. ESTHER, the database of the α/β -hydrolase fold superfamily of proteins: Tools to explore diversity of functions. *Nucleic Acids Research*, 41(D1):D423–9, 2013.
- [134] Jean Louis ARPIGNY and Karl-Erich JAEGER. Bacterial lipolytic enzymes: classification and properties. *Biochemical Journal*, 343(1):177, 1999.
- [135] Dimitra Zarafeta, Danai Moschidi, Efthymios Ladoukakis, Sergey Gavrilov, Evangelia D. Chrysina, Aristotelis Chatziioannou, Ilya Kublanov, Georgios Skretas, and Fragiskos N. Kolisis. Metagenomic mining for thermostable esterolytic enzymes uncovers a new family of bacterial esterases. *Scientific Reports*, 6(1):38886, 2016.
- [136] Mariana Rangel Pereira. *Prospecção de genes codificadores de enzimas lipolíticas em biblioteca metageômica de consórcio microbiano degradador de óleo diesel*. Doctoral thesis, Universidade de São Paulo, São Paulo, 2011.
- [137] Johan Peränen, Marja Rikonen, Marko Hyvönen, and Leevi Kääriäinen. T7 Vectors with a Modified T7/ac Promoter for Expression of Proteins in *Escherichia coli*. *Analytical Biochemistry*, 236(2):371–373, 1996.
- [138] Monique M. Martin and Lars Lindqvist. The pH dependence of fluorescein fluorescence. *Journal of Luminescence*, 10(6):381–390, 1975.
- [139] Joseph Zock, Cathleen Cantwell, James Swartling, Roland Hodges, Tonya Pohl, Kimberly Sutton, Paul Rosteck, Derek McGilvray, and Stephen Queener. The *Bacillus subtilis* pnbA gene encoding p-nitrobenzyl esterase: cloning, sequence and high-level expression in *Escherichia coli*. *Gene*, 151(1-2):37–43, 1994.
- [140] Susanne Wilhelm, Jan Tommassen, and Karl Erich Jaeger. A novel lipolytic enzyme located in the outer membrane of *Pseudomonas aeruginosa*. *Journal of Bacteriology*, 181(22):6977–6986, 1999.

- [141] A Raibaud, M Zalacain, T G Holt, R Tizard, and C J Thompson. Nucleotide sequence analysis reveals linked regulatory genes encoded by the bialaphos biosynthetic gene cluster of *Streptomyces* Nucleotide Sequence Analysis Reveals Linked N-Acetyl Hydrolase, Thioesterase, Transport, and Regulatory Genes Encoded by the. *J. Bacteriol.*, 173(14):4454–63, 1991.
- [142] E. Luthi, D. R. Love, J. McAnulty, C. Wallace, P. A. Caughey, D. Saul, and P. L. Bergquist. Cloning, sequence analysis, and expression of genes encoding Xylan-degrading enzymes from the thermophile *Caldocellum saccharolyticum*. *Applied and Environmental Microbiology*, 56(4):1017–1024, 1990.
- [143] J Anguita, L B Rodríguez Aparicio, and G Naharro. Purification, gene cloning, amino acid sequence analysis, and expression of an extracellular lipase from an *Aeromonas hydrophila* human isolate. *Applied and Environmental Microbiology*, 59(8):2411–7, 1993.
- [144] C.J. Duggleby and P.A. Williams. Purification and some properties of the 2-hydroxy-6-oxohepta-2,4-dienoate hydrolase (2-hydroxymuconic semialdehyde hydrolase) encoded by the TOL plasmid pWW0 from *Pseudomonas putida* mt-2. *Journal of General Microbiology*, 132(3):717–726, 1986.
- [145] Young Jun Park, Sung Jin Yoon, and Hee Bong Lee. A novel thermostable arylesterase from the archaeon *Sulfolobus solfataricus* P1: Purification, characterization, and expression. *Journal of Bacteriology*, 190(24):8086–8095, 2008.
- [146] Myung Hwan Lee, Kyung Sik Hong, Shweta Malhotra, Ji Hye Park, Eul Chul Hwang, Hong Kyu Choi, Young Sup Kim, Weixin Tao, and Seon Woo Lee. A new esterase EstD2 isolated from plant rhizosphere soil metagenome. *Applied Microbiology and Biotechnology*, 88(5):1125–1134, 2010.
- [147] Min Keun Kim, Tae Ho Kang, Jung-ho Kim, Hoon Kim, and Han Dae Yun. Cloning and identification of a new group esterase (Est5s) from noncultured rumen bacterium. *Journal of Microbiology and Biotechnology*, 22(8):1044–1053, 2012.
- [148] B. van den Berg. Crystal Structure of a Full-Length Autotransporter. *Journal of Molecular Biology*, 396(3):627–633, 2010.
- [149] Chen Li, Jian J. Li, Mark G. Montgomery, Stephen P. Wood, and T. D H Bugg. Catalytic role for arginine 188 in the C-C hydrolase catalytic mechanism for *Escherichia coli* MhpC and *Burkholderia xenovorans* LB400 BphD. *Biochemistry*, 45(41):12470–12479, 2006.
- [150] Nathan A. Lack, Katherine C. Yam, Edward E. Lowe, Geoff P. Horsman, Robin L. Owen, Edith Sim, and Lindsay D. Eltis. Characterization of a carbon-carbon hydrolase from *Mycobacterium tuberculosis* involved in cholesterol metabolism. *Journal of Biological Chemistry*, 285(1):434–443, 2010.
- [151] Chen Li, Melanie Hassler, and T. D H Bugg. Catalytic promiscuity in the α/β -hydrolase superfamily: Hydroxamic acid formation, C-C bond formation, ester and thioester hydrolysis in the C-C hydrolase family. *ChemBioChem*, 9(1):71–76, 2008.

- [152] María Alcaide, Jesús Tornés, Peter J. Stogios, Xiaohui Xu, Christoph Gertler, Rosa Di Leo, Rafael Bargiela, Álvaro Lafraya, María-Eugenia Guazzaroni, Nieves López-Cortés, Tatyana N. Chernikova, Olga V. Golyshina, Taras Y. Nechitaylo, Iris Plumeier, Dietmar H. Pieper, Michail M. Yakimov, Alexei Savchenko, Peter N. Golyshin, and Manuel Ferrer. Single residues dictate the co-evolution of dual esterases: MCP hydrolases from the α/β hydrolase family. *Biochemical Journal*, 454(1):157–166, 2013.
- [153] Holly J. Atkinson, John H. Morris, Thomas E. Ferrin, and Patricia C. Babbitt. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE*, 4(2):e4345, 2009.
- [154] L. David, Eong Cheah, Mirosław Cygler, Bauke Dijkstra, Felix Frolov, M. Sybille, Michal Harel, S. James Remington, Israel Silman, Joseph Schrag, L. Joel, H. G. Verschueren Koen, and Adrian Goldman. The α/β hydrolase fold. *Protein Engineering, Design and Selection*, 5(3):197–211, 1992.
- [155] David L Ollis and Paul D Carr. α/β Hydrolase Fold: An Update. *Protein & Peptide Letters*, 16(10):1137–1148, 2009.
- [156] Anthony J Kirby and Florian Hollfelder. *From Enzyme Models to Model Enzymes*. Royal Society of Chemistry, Cambridge, 2009.
- [157] K-E. Jaeger, B. W. Dijkstra, and M. T. Reetz. Bacterial Biocatalysts: Molecular Biology, Three-Dimensional Structures, and Biotechnological Applications of Lipases. *Annual Review of Microbiology*, 53(1):315–351, 1999.
- [158] Arren Bar-Even, Elad Noor, Yonatan Savir, Wolfram Liebermeister, Dan Davidi, Dan S. Tawfik, and Ron Milo. The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, 50(21):4402–4410, 2011.
- [159] Kathy Huynh and Carrie L. Partch. Analysis of protein stability and ligand interactions by thermal shift assay. *Current Protocols in Protein Science*, 2015:19.26.1–19.26.14, 2015.
- [160] T. Yamada, J. Nakamura, S. Kawamura, and K. Tanaka. Hexagonal overlayer of hydroxyl groups on Ni(100) modified by sulfur: effect on the H₂-D₂ isotope exchange reaction. *Journal of Electron Spectroscopy and Related Phenomena*, 44(1):79–88, 1987.
- [161] Fabrizio Pucci and Marianne Rومان. Physical and molecular bases of protein thermal stability and cold adaptation, 2017.
- [162] Evamaria I. Petersen, Goran Valinger, Beate Sölkner, Gerhard Stubenrauch, and Helmut Schwab. A novel esterase from *Burkholderia gladioli* which shows high deacetylation activity on cephalosporins is related to β -lactamases and DD-peptidases. *Journal of Biotechnology*, 89(1):11–25, 2001.
- [163] Dolores Pérez, Filip Kovačić, Susanne Wilhelm, Karl Erich Jaeger, María Teresa García, Antonio Ventosa, and Encarnación Mellado. Identification of amino acids involved in the hydrolytic activity of lipase LipBL from *Marinobacter lipolyticus*. *Microbiology (United Kingdom)*, 158(8):2192–2203, 2012.

- [164] Konanani Rashamuse, Victoria Magomani, Tina Ronneburg, and Dean Brady. A novel family VIII carboxylesterase derived from a leachate metagenome library exhibits promiscuous β -lactamase activity on nitrocefin. *Applied Microbiology and Biotechnology*, 83(3):491–500, 2009.
- [165] Sun Shin Cha, Young Jun An, Chang Sook Jeong, Min Kyu Kim, Jeong Ho Jeon, Chang Muk Lee, Hyun Sook Lee, Sung Gyun Kang, and Jung Hyun Lee. Structural basis for the β -lactamase activity of EstU1, a family VIII carboxylesterase. *Proteins: Structure, Function and Bioinformatics*, 81(11):2045–2051, 2013.
- [166] Mark Jones and Michael I Page. An Esterase with b-Lactamase Activity. *Journal of the Chemical Society, Chemical Communications*, 0(316):316–317, 1991.
- [167] B. van Loo, S. Jonas, A. C. Babbie, A. Benjdia, O. Berteau, M. Hyvonen, and F. Hollfelder. An efficient, multiply promiscuous hydrolase in the alkaline phosphatase superfamily. *Proceedings of the National Academy of Sciences*, 107(7):2740–2745, 2010.
- [168] Anthony L Schillmiller, Karin Gilgallon, Banibrata Ghosh, A Daniel Jones, and Robert L Last. Acylsugar Acylhydrolases: Carboxylesterase-Catalyzed Hydrolysis of Acylsugars in Tomato Trichomes. *Plant physiology*, 170(3):1331–44, 2016.
- [169] F. Hollfelder, A. J. Kirby, and D. S. Tawfik. On the magnitude and specificity of medium effects in enzyme-like catalysts for proton transfer. *Journal of Organic Chemistry*, 66(17):5866–5874, 2001.
- [170] Vandana Lamba, Enis Sanchez, Lauren Rose Fanning, Kathryn Howe, Maria Alejandra Alvarez, Daniel Herschlag, and Marcello Forconi. Kemp eliminase activity of ketosteroid isomerase. *Biochemistry*, 56(4):582–591, 2017.
- [171] Claudia Schmidt-Dannert and Frances H. Arnold. Directed evolution of industrial enzymes. *Trends in Biotechnology*, 17(4):135–136, 1999.
- [172] Yu Yang, Jun Yang, and Lei Jiang. Comment on "a bacterium that degrades and assimilates poly(ethylene terephthalate)". *Science*, 353(6301):759, 2016.
- [173] U. Bergthorsson, D. I. Andersson, and J. R. Roth. Ohno's dilemma: Evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences*, 104(43):17004–17009, 2007.
- [174] Shelley D Copley. Toward a systems biology perspective on enzyme evolution. *The Journal of biological chemistry*, 287(1):3–10, 2012.
- [175] Klaus Ruthenberg. Schrader, Paul Gerhard Heinrich. In *Neue Deutsche Biographie, Band 23*, page 508. 2007.
- [176] N. Sethunathan and T. Yoshida. A *Flavobacterium* sp. that degrades diazinon and parathion. *Canadian Journal of Microbiology*, 19(7):873–875, 1973.
- [177] D M Munnecke and D P Hsieh. Microbial decontamination of parathion and p-nitrophenol in aqueous media. *Applied microbiology*, 28(2):212–217, 1974.

- [178] Steven R. Caldwell, Jennifer R. Newcomb, Kristina A. Schlecht, and Frank M. Raushel. Limits of diffusion in the hydrolysis of substrates by the phosphotriesterase from *Pseudomonas diminuta*. *Biochemistry*, 30(30):7438–7444, 1991.
- [179] Livnat Afriat, Cintia Roodveldt, Giuseppe Manco, and Dan S. Tawfik. The latent promiscuity of newly identified microbial lactonases is linked to a recently diverged phosphotriesterase. *Biochemistry*, 45(46):13677–13686, 2006.
- [180] Livnat Afriat-Jurnou, Colin J. Jackson, and Dan S. Tawfik. Reconstructing a missing link in the evolution of a recently diverged phosphotriesterase by active-site loop remodeling. *Biochemistry*, 51(31):6047–6055, 2012.
- [181] D. S. Kemp, Daniel D. Cox, and Kenneth G. Paul. The Physical Organic Chemistry of Benzisoxazoles. IV. The Origins and Catalytic Nature of the Solvent Rate Acceleration for the Decarboxylation of 3-Carboxybenzisoxazoles. *Journal of the American Chemical Society*, 97(25):7312–7318, 1975.
- [182] Paul G. McCracken, Colin G. Ferguson, Dragos Vizitiu, Caroline S. Walkinshaw, Yu Wang, and Gregory R.J. Thatcher. Amino-cyclodextrins as biomimetics: Catalysis of the Kemp elimination. *Journal of the Chemical Society. Perkin Transactions 2*, (5):911–912, 1999.
- [183] Alan J. Kennan and H. W. Whitlock. Host-Catalyzed Isoxazole Ring Opening: A Rationally Designed Artificial Enzyme. *Journal of the American Chemical Society*, 118(12):3027–3028, 1996.
- [184] William Cullen, M. Cristina Misuraca, Christopher A. Hunter, Nicholas H. Williams, and Michael D. Ward. Highly efficient catalysis of the Kemp elimination in the cavity of a cubic coordination cage. *Nature Chemistry*, 8(3):231–236, 2016.
- [185] Jorge Pérez-Juste, Florian Hollfelder, Anthony J. Kirby, and Jan B.F.N. Engberts. Vesicles accelerate proton transfer from carbon up to 850-fold. *Organic Letters*, 2(2):127–130, 2000.
- [186] Enis Sanchez, Saoran Lu, Carson Reed, Joshua Schmidt, and Marcello Forconi. Kemp elimination in cationic micelles: Designed enzyme-like rates achieved through the addition of long-chain bases. *Journal of Physical Organic Chemistry*, 29(4):185–189, 2016.
- [187] F. Hollfelder, A. J. Kirby, and D. S. Tawfik. Efficient catalysis of proton transfer by synzymes. *Journal of the American Chemical Society*, 119(40):9578–9579, 1997.
- [188] Donald Hilvert. Critical Analysis of Antibody Catalysis. *Annual Review of Biochemistry*, 69(1):751–793, 2000.
- [189] Simon N. Thorn, Richard G. Daniels, Maria Teresa M. Auditor, and Donald Hilvert. Large rate accelerations in antibody catalysis by strategic use of haptenic charge. *Nature*, 373(6511):228–230, 1995.

- [190] Arnaud Genre-Grandpierre, Charles Tellier, Marie Jeanne Loirat, Dominique Blanchard, David R.W. Hodgson, Florian Hollfelder, and Anthony J. Kirby. Catalysis of the Kemp elimination by antibodies elicited against a cationic hapten. *Bioorganic and Medicinal Chemistry Letters*, 7(19):2497–2502, 1997.
- [191] Kazuya Kikuchi, Renate B. Hannak, Mao Jun Guo, Anthony J. Kirby, and Donald Hilvert. Toward bifunctional antibody catalysis. *Bioorganic and Medicinal Chemistry*, 14(18):6189–6196, 2006.
- [192] Florian P. Seebeck and Donald Hilvert. Positional ordering of reacting groups contributes significantly to the efficiency of proton transfer at an antibody active site. *Journal of the American Chemical Society*, 127(4):1307–1312, 2005.
- [193] Florian Hollfelder, Anthony J. Kirby, and Dan S. Tawfik. Off-the-shelf proteins that rival tailor-made antibodies as catalysts. *Nature*, 383(6595):60–63, 1996.
- [194] M. Merski and B. K. Shoichet. Engineering a model protein cavity to catalyze the Kemp elimination. *Proceedings of the National Academy of Sciences*, 109(40):16179–16183, 2012.
- [195] I. V. Korendovych, D. W. Kulp, Y. Wu, H. Cheng, H. Roder, and W. F. DeGrado. Design of a switchable eliminase. *Proceedings of the National Academy of Sciences*, 108(17):6823–6827, 2011.
- [196] Aitao Li, Binju Wang, Adriana Ilie, Kshatresh D. Dubey, Gert Bange, Ivan V. Korendovych, Sason Shaik, and Manfred T. Reetz. A redox-mediated Kemp eliminase. *Nature Communications*, 8:14876, 2017.
- [197] Olga Khersonsky, Daniela Röthlisberger, Orly Dym, Shira Albeck, Colin J. Jackson, David Baker, and Dan S. Tawfik. Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the ke07 series. *Journal of Molecular Biology*, 396(4):1025–1042, 2010.
- [198] Olga Khersonsky, Daniela Röthlisberger, Andrew M. Wollacott, Paul Murphy, Orly Dym, Shira Albeck, Gert Kiss, K. N. Houk, David Baker, and Dan S. Tawfik. Optimization of the in-silico-designed Kemp eliminase KE70 by computational design and directed evolution. *Journal of Molecular Biology*, 407(3):391–412, 2011.
- [199] O. Khersonsky, G. Kiss, D. Rothlisberger, O. Dym, S. Albeck, K. N. Houk, D. Baker, and D. S. Tawfik. Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proceedings of the National Academy of Sciences*, 109(26):10358–10363, 2012.
- [200] Yurii S. Moroz, Tiffany T. Dunston, Olga V. Makhlynets, Olesia V. Moroz, Yibing Wu, Jennifer H. Yoon, Alissa B. Olsen, Jaclyn M. McLaughlin, Korrie L. Mack, Pallavi M. Gosavi, Nico A.J. Van Nuland, and Ivan V. Korendovych. New Tricks for Old Proteins: Single Mutations in a Nonenzymatic Protein Give Rise to Various Enzymatic Activities. *Journal of the American Chemical Society*, 137(47):14905–14911, 2015.
- [201] Yufeng Miao, Richard Metzner, and Yasuhisa Asano. Kemp Elimination Catalyzed by Naturally Occurring Aldoxime Dehydratases. *ChemBioChem*, 18(5):451–454, 2017.

- [202] Ivan V. Korendovych and William F. DeGrado. Catalytic efficiency of designed catalytic proteins. *Current Opinion in Structural Biology*, 27(1):113–121, 2014.
- [203] Kazuya Kikuchi, Simon N. Thorn, and Donald Hilvert. Albumin-catalyzed proton transfer. *Journal of the American Chemical Society*, 118(34):8184–8185, 1996.
- [204] Florian Hollfelder, Anthony J. Kirby, Dan S. Tawfik, Kazuya Kikuchi, and Donald Hilvert. Characterization of proton-transfer catalysis by serum albumins. *Journal of the American Chemical Society*, 122(6):1022–1029, 2000.
- [205] Olga Khersonsky, Sergey Malitsky, Ilana Rogachev, and Dan S. Tawfik. Role of chemistry versus substrate binding in recruiting promiscuous enzyme functions. *Biochemistry*, 50(13):2683–2690, 2011.
- [206] Masanari Kitagawa, Takeshi Ara, Mohammad Arifuzzaman, Tomoko Ioka-Nakamichi, Eiji Inamoto, Hiromi Toyonaga, and Hirotada Mori. Complete set of ORF clones of Escherichia coli ASKA library (A complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Research*, 12(5):291–299, 2005.
- [207] Youssr Skhiri, Philipp Gruner, Benoît Semin, Quentin Brosseau, Deniz Pekin, Linas Mazutis, Victoire Goust, Felix Kleinschmidt, Abdeslam El Harrak, J. Brian Hutchison, Estelle Mayot, Jean François Bartolo, Andrew D. Griffiths, Valérie Taly, and Jean Christophe Baret. Dynamics of molecular transport by surfactants in emulsions. *Soft Matter*, 8(41):10618–10627, 2012.
- [208] Lucas Frenz, Kerstin Blank, Eric Brouzes, and Andrew D. Griffiths. Reliable microfluidic on-chip incubation of droplets in delay-lines. *Lab on a Chip*, 9(10):1344–1348, 2009.
- [209] Cyrus Chothia, Julian Gough, Christine Vogel, and Sarah A. Teichmann. Evolution of the protein repertoire. *Science*, 300(5626):1701–1703, 2003.
- [210] Jian Qun Chen, Ying Wu, Haiwang Yang, Joy Bergelson, Martin Kreitman, and Dacheng Tian. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Molecular Biology and Evolution*, 26(7):1523–1531, 2009.
- [211] Ágnes Tóth-Petróczy and Dan S. Tawfik. Protein insertions and deletions enabled by neutral roaming in sequence space. *Molecular Biology and Evolution*, 30(4):761–771, 2013.
- [212] Romain A. Studer, Benoit H. Dessailly, and Christine A. Orengo. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochemical Journal*, 449(3):581–594, 2013.
- [213] Jeffrey I Boucher, Joseph R Jacobowitz, Brian C Beckett, Scott Classen, and Douglas L Theobald. An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases. *eLife*, 3, 2014.

- [214] Denis Odokonyero, Ayano Sakai, Yury Patskovsky, Vladimir N Malashkevich, A. A. Fedorov, Jeffrey B Bonanno, Elena V Fedorov, Rafael Toro, Rakhi Agarwal, Chenxi Wang, Nicole D S Ozerova, Wen Shan Yew, J Michael Sauder, Subramanyam Swaminathan, Stephen K Burley, Steven C Almo, and Margaret E Glasner. Loss of quaternary structure is associated with rapid sequence divergence in the OSBS family. *Proceedings of the National Academy of Sciences*, 111(23):8535–8540, 2014.
- [215] Yakov Kipnis, Eynat Dellus-Gur, and Dan S. Tawfik. TRINS: A method for gene modification by randomized tandem repeat insertions. *Protein Engineering, Design and Selection*, 25(9):437–444, 2012.
- [216] R. Craig Cadwell and Gerald F. Joyce. Randomization of genes by PCR mutagenesis. *Genome Research*, 2(1):28–33, 1992.
- [217] Manfred T. Reetz. Gene Mutagenesis Methods. In *Directed Evolution of Selective Enzymes*, pages 59–114. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2016.
- [218] Andrew E. Firth and Wayne M. Patrick. GLUE-IT and PEDEL-AA: new programmes for analyzing protein diversity in randomized libraries. *Nucleic acids research*, 36(Web Server issue):W281–W285, 2008.
- [219] Matteo Paolo Ferla. Mutanalyst, an online tool for assessing the mutational spectrum of epPCR libraries with poor sampling. *BMC Bioinformatics*, 17(1):152, 2016.
- [220] Grégory Boël, Reka Letso, Helen Neely, W Nicholson Price, Kam Ho Wong, Min Su, Jon D. Luff, Mayank Valecha, John K Everett, Thomas B Acton, Rong Xiao, Gaetano T Montelione, Daniel P Aalberts, and John F Hunt. Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature*, 529(7586):358–363, 2016.
- [221] Amir Aharoni, Leonid Gaidukov, Shai Yagur, Lilly Toker, Israel Silman, and Dan S Tawfik. Directed evolution of mammalian paraoxonases PON1 and PON3 for bacterial expression and catalytic specialization. *Proceedings of the National Academy of Sciences*, 101(2):482–487, 2004.
- [222] Raymond D. Socha and Nobuhiko Tokuriki. Modulating protein stability - Directed evolution strategies for improved protein function, 2013.
- [223] Ryang Guk Kim and Jun Tao Guo. Systematic analysis of short internal indels and their impact on protein folding. *BMC Structural Biology*, 10(1):24, 2010.
- [224] Courtney E. Gonzalez, Paul Roberts, and Marc Ostermeier. Fitness Effects of Single Amino Acid Insertions and Deletions in TEM-1 β -Lactamase. *Journal of Molecular Biology*, 431(12):2320–2330, 2019.
- [225] Oliviero Carugo. How large B-factors can be in protein crystal structures. *BMC Bioinformatics*, 19(1):61, 2018.
- [226] Zhoutong Sun, Qian Liu, Ge Qu, Yan Feng, and Manfred T. Reetz. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chemical Reviews*, 119(3):1626–1665, 2019.

- [227] Evgeny V. Leushkin, Georgii A. Bazykin, and Alexey S. Kondrashov. Insertions and deletions trigger adaptive walks in *Drosophila* proteins. *Proceedings of the Royal Society B: Biological Sciences*, 279(1740):3075–3082, 2012.
- [228] Philip A. Romero and Frances H. Arnold. Exploring protein fitness landscapes by directed evolution, 2009.
- [229] Shingo Sakamoto, Toru Komatsu, Tasuku Ueno, Kenjiro Hanaoka, and Yasuteru Urano. Fluorescence detection of serum albumin with a turnover-based sensor utilizing Kemp elimination reaction. *Bioorganic and Medicinal Chemistry Letters*, 27(15):3464–3467, 2017.
- [230] Bernard Valeur. Characteristics of Fluorescence Emission. In *Molecular Fluorescence*, pages 53–74. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2012.
- [231] Gabrielle Woronoff, Abdeslam El Harrak, Estelle Mayot, Olivier Schicke, Oliver J. Miller, Patrice Soumillion, Andrew D. Griffiths, and Michael Ryckelynck. New Generation of Amino Coumarin Methyl Sulfonate-Based Fluorogenic Substrates for Amidase Assays in Droplet-Based Microfluidic Applications. *Analytical Chemistry*, 83(8):2852–2857, 2011.
- [232] Zachary P. Demko and K. Barry Sharpless. A Click Chemistry Approach to Tetrazoles by Huisgen 1,3-Dipolar Cycloaddition: Synthesis of 5-Acyltetrazoles from Azides and Acyl Cyanides We thank the National Institute of General Medical Sciences, National Institutes of Health (GM-28384), the National . *Angewandte Chemie International Edition*, 41(12):2113, 2002.
- [233] David Amantini, Romina Beleggia, Francesco Fringuelli, Ferdinando Pizzo, and Luigi Vaccaro. TBAF-Catalyzed Synthesis of 5-Substituted 1 H -Tetrazoles under Solventless Conditions. *The Journal of Organic Chemistry*, 69(8):2896–2898, 2004.
- [234] Sameer M. Joshi, Rasika B. Mane, Krishna R. Pulagam, Vanessa Gomez-Vallejo, Jordi Llop, and Chandrashekhar Rode. The microwave-assisted synthesis of 5-substituted 1: H -tetrazoles via [3+2] cycloaddition over a heterogeneous Cu-based catalyst: Application to the preparation of ¹³N-labelled tetrazoles. *New Journal of Chemistry*, 41(16):8084–8091, 2017.
- [235] J. P. Ferris and F. R. Antonucci. Mechanisms of the Photochemical Rearrangements of Ortho-Substituted Benzene Derivatives and Related Heterocycles. *Journal of the American Chemical Society*, 96(7):2014–2019, 1974.
- [236] Cláudio M. Nunes, Sandra M.V. Pinto, Igor Reva, and Rui Fausto. On the Photochemistry of 1,2-Benzisoxazole: Capture of Elusive Spiro-2H-azirine and Ketenimine Intermediates. *European Journal of Organic Chemistry*, 2016(24):4152–4158, 2016.
- [237] G T Hermanson. Heterobifunctional Crosslinkers. In *Bioconjugate techniques*, pages 276–335. Academic Press, 2008.
- [238] Celia Lee Go, Walter H. Waddell, and Walter H. Waddell. Evolution of Photooxidation Products upon Irradiation of Phenyl Azide in the Presence of Molecular Oxygen. *Journal of Organic Chemistry*, 48(17):2897–2900, 1983.

- [239] Nina Gritsan and Matthew Platz. Photochemistry of Azides: The Azide/Nitrene Interface. In *Organic Azides: Syntheses and Applications*, pages 311–372. 2011.
- [240] Samuel J. Lord, Nicholas R. Conley, Hsiao Lu D. Lee, Reichel Samuel, Na Liu, Robert J. Twieg, and W. E. Moerner. A photoactivatable push-pull fluorophore for single-molecule imaging in live cells. *Journal of the American Chemical Society*, 130(29):9204–9205, 2008.
- [241] Annapaola Migani, Verónica Leyva, Ferran Feixas, Thomas Schmierer, Peter Gilch, Inés Corral, Leticia González, and Lluís Blancafort. Ultrafast irreversible phototautomerization of o-nitrobenzaldehyde. *Chemical Communications*, 47(22):6383–6385, 2011.
- [242] Jianzhang Zhao, Shaomin Ji, Yinghui Chen, Huimin Guo, and Pei Yang. Excited state intramolecular proton transfer (ESIPT): From principal photophysics to the development of new chromophores and applications in fluorescent molecular probes and luminescent materials, 2012.
- [243] Joseph R. Lakowicz. Solvent and Environmental Effects. In *Principles of Fluorescence Spectroscopy*, chapter 6, pages 205–235. Springer, Singapore, third edit edition, 2006.
- [244] Kathleen R. Murphy. A note on determining the extent of the water Raman peak in fluorescence spectroscopy. *Applied Spectroscopy*, 65(2):233–236, 2011.
- [245] Mortimer J. Kamlet, José Luis M. Abboud, Michael H. Abraham, and R. W. Taft. Linear Solvation Energy Relationships. 23. A Comprehensive Collection of the Solvatochromic Parameters, π , α , and β , and Some Methods for Simplifying the Generalized Solvatochromic Equation. *Journal of Organic Chemistry*, 48(17):2877–2887, 1983.
- [246] Esther M. Gabor, Erik J. De Vries, and Dick B. Janssen. Construction, characterization, and use of small-insert gene banks of DNA isolated from soil and enrichment cultures for the recovery of novel amidases. *Environmental Microbiology*, 6(9):948–958, 2004.
- [247] Douglas Hanahan. Studies on transformation of *Escherichia coli* with plasmids. *Journal of Molecular Biology*, 166(4):557–580, 1983.
- [248] D J Catanese, J M Fogg, D E Schrock, B E Gilbert, and L Zechiedrich. Supercoiled Minivector DNA resists shear forces associated with gene therapy delivery. *Gene Therapy*, 19(1):94–100, 2012.
- [249] Barbara Röder, Karin Frühwirth, Claus Vogl, Martin Wagner, and Peter Rossmannith. Impact of long-term storage on stability of standard DNA for nucleic acid-based methods. *Journal of Clinical Microbiology*, 48(11):4260–4262, 2010.
- [250] D. Travis Gallagher, Natasha N. Smith, Sook-Kyung Kim, Annie Heroux, Howard Robinson, and Prasad T. Reddy. Structure of the Class IV Adenylyl Cyclase Reveals a Novel Fold. *Journal of Molecular Biology*, 362(1):114–122, 2006.
- [251] D. Travis Gallagher, Sook Kyung Kim, Howard Robinson, and Prasad T. Reddy. Active-site structure of class IV adenylyl cyclase and transphyletic mechanism. *Journal of Molecular Biology*, 405(3):787–803, 2011.

- [252] Lawrence A Kelley, Stefans Mezulis, Christopher M Yates, Mark N Wass, and Michael J E Sternberg. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, 10(6):845–58, 2015.
- [253] Joseph R Lakowicz, editor. *Fluorescence Sensing*, pages 623–673. Springer US, Boston, MA, 2006.
- [254] John McCafferty, Andrew D. Griffiths, Greg Winter, and David J. Chiswell. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature*, 348(6301):552–554, 1990.
- [255] Tim Clackson, Hennie R. Hoogenboom, Andrew D. Griffiths, and Greg Winter. Making antibody fragments using phage display libraries. *Nature*, 352(6336):624–628, 1991.
- [256] Nobuhiko Tokuriki, Colin J. Jackson, Livnat Afriat-Jurnou, Kirsten T. Wyganowski, Renmei Tang, and Dan S. Tawfik. Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme. *Nature Communications*, 3(1):1257, 2012.
- [257] Charlotte M Miton, Stefanie Jonas, Gerhard Fischer, Fernanda Duarte, Mark F Mohamed, Bert van Loo, Bálint Kintszes, Shina C L Kamerlin, Nobuhiko Tokuriki, Marko Hyvönen, and Florian Hollfelder. Evolutionary repurposing of a sulfatase: A new Michaelis complex leads to efficient transition state charge offset. *Proceedings of the National Academy of Sciences*, 115(31):201607817, 2018.
- [258] C Beloin, A Roux, and J M Ghigo. Escherichia coli biofilms. *Current Topics in Microbiology and Immunology*, 322:249–289, 2008.
- [259] Harley H McAdams and Adam Arkin. It’s a noisy business! Genetic regulation at the nanomolar scale, 1999.
- [260] Maria L. Kilfoil, Paul Lasko, and Ehab Abouheif. Stochastic variation: From single cells to superorganisms. *HFSP Journal*, 3(6):379–385, 2009.
- [261] EM Gabor. *Harvesting novel biocatalysts from the metagenome*. PhD thesis, University of Groningen, 2004.
- [262] Esther M. Gabor, Erik J. De Vries, and Dick B. Janssen. Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. *FEMS Microbiology Ecology*, 44(2):153–163, 2003.
- [263] Peter Eyer, Franz Worek, Daniela Kiderlen, Goran Sinko, Anita Stuglin, Vera Simeon-Rudolf, and Elsa Reiner. Molar absorption coefficients for the reduced ellman reagent: Reassessment. *Analytical Biochemistry*, 312(2):224–227, 2003.
- [264] Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, 2010.
- [265] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009.

-
- [266] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–504, 2003.
- [267] Jason J. Nichols, P. Ewen King-Smith, Erich A. Hinel, Miru Thangavelu, and Kelly K. Nichols. The use of fluorescent quenching in studying the contribution of evaporation to tear thinning. *Investigative Ophthalmology and Visual Science*, 53(9):5426–5432, 2012.

Appendix A

Supplementary Data Chapter 2

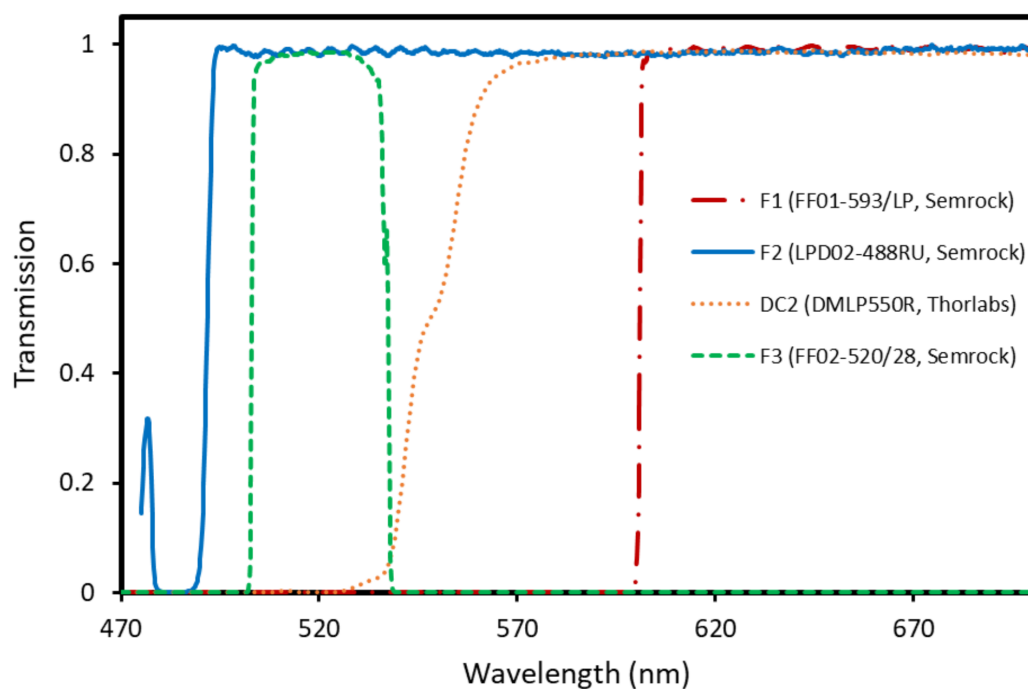


Fig. A.1 Transmission spectra of the filters used in the FADS.

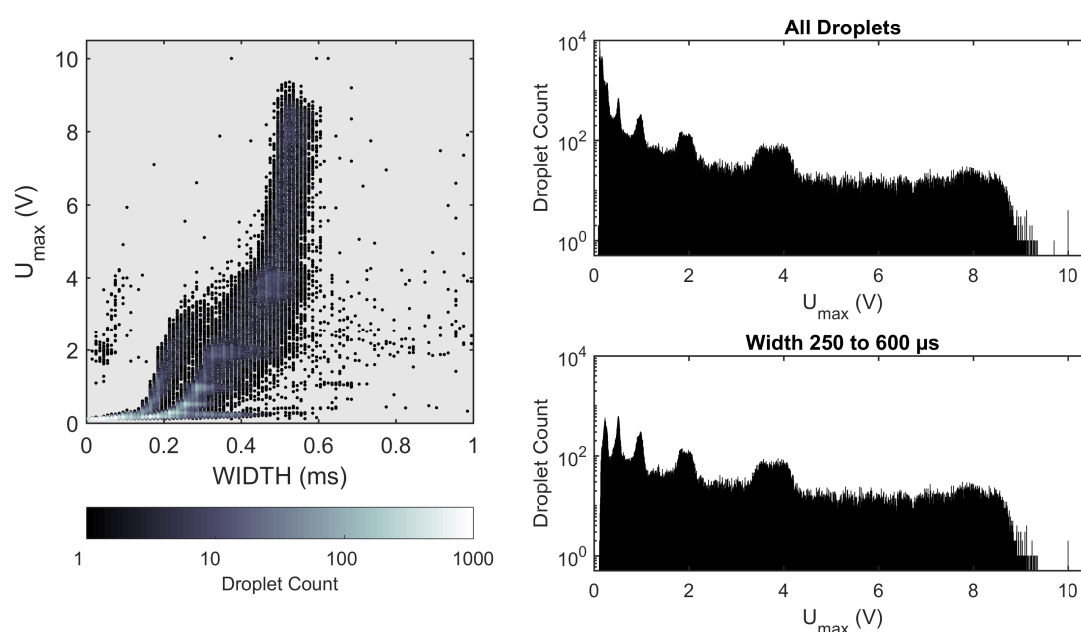


Fig. A.2 Shown is the raw data from which the linearity of fluorescence measurements in Figure 2.8 was derived. The peaks were detected using peak detection in Matlab and assigned concentrations from the highest to the lowest peak, the width of peaks was measured at half peak prominence.

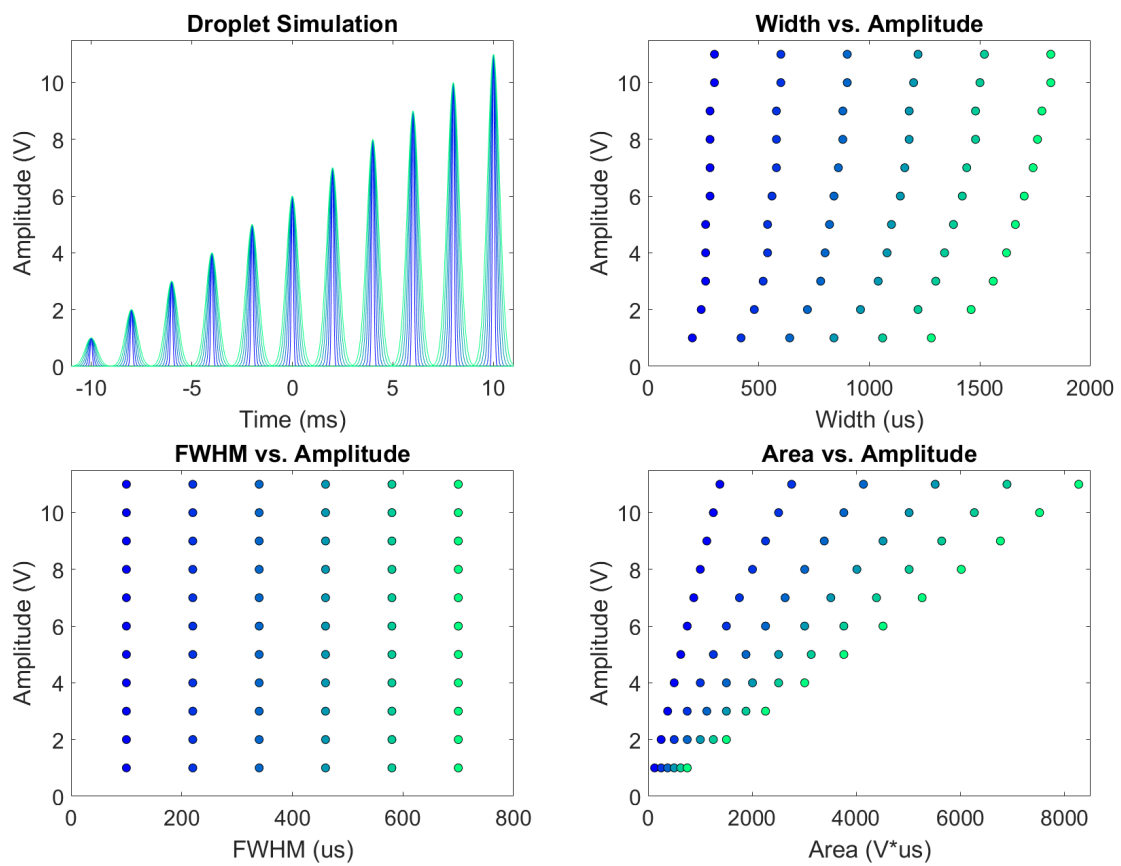


Fig. A.3 Shown is a simulation of different fluorescence signals based on normal distributions with various amplitudes and standard deviations (relating to droplet size). The width, FWHM, and area were then calculated according to the algorithm in Figure 2.4.

Appendix B

Supplementary Data Chapter 3

Table B.1 The SCV library is a combination of several metagenomic libraries in plasmid pZero2 [67]. The average library member is 3 to 5 kbp in size.

Library	Library Size /10 ³	Source
ENR-M	23	Marine Sludge
ENR-S	35	Goose Pond
ENR-G	25	Sandy Soil
ENR-L	30	Loamy Soil
DIR-L	80	Loamy Soil
SEM	80	Vanilla Pods
TSA	45	Vanilla Pods
DIR-MC	500	Medium Compost
DIR-RC	300	Thermophilic Compost
CR2	135	Cow Rumen
SCV	1,253	

B.1 DNA Inserts

B.1.1 N1 DNA Insert

```
1  tgcgaaaaggaaaatatccggctgggcagcgtggaaggaatcggggcggcagatcatgcggtcacggcctgtatgatgtggcgaaaccgc
91  cagtaccataaaaaagagctgaacggcccatggaaatttctccctgctgggcactgttacaagaaaagacggcgcagtgatatctgcat
181  ctgcacatcaacctgtgcaatacggaaatgcagatcctgggcggtcatctgaacgaatgccgcatcggcgccaccggtgaaattattgtg
271  cggacgattccgggacaggtaggccgtctgctggtcgacaaggtgaccgggctgaacctgtttcagtttgaagggtaacggagtgtgttc
361  gctcttttctaaccactaaccgggaggaactatgaaccagcaaaccatatatcgcatcgttaaattttcagaagaccgggactgggat
451  cagttccatacggcgccaatctggccaagtctatttccattgaggccaacgaactgctggagtgttttagtggagcgatacggagtat
541  gatgtgtcccggtccgggaagagctggccgatgtgctggtctattgccggaatatgctggacaggctggaactggatgaagatgagatc
```

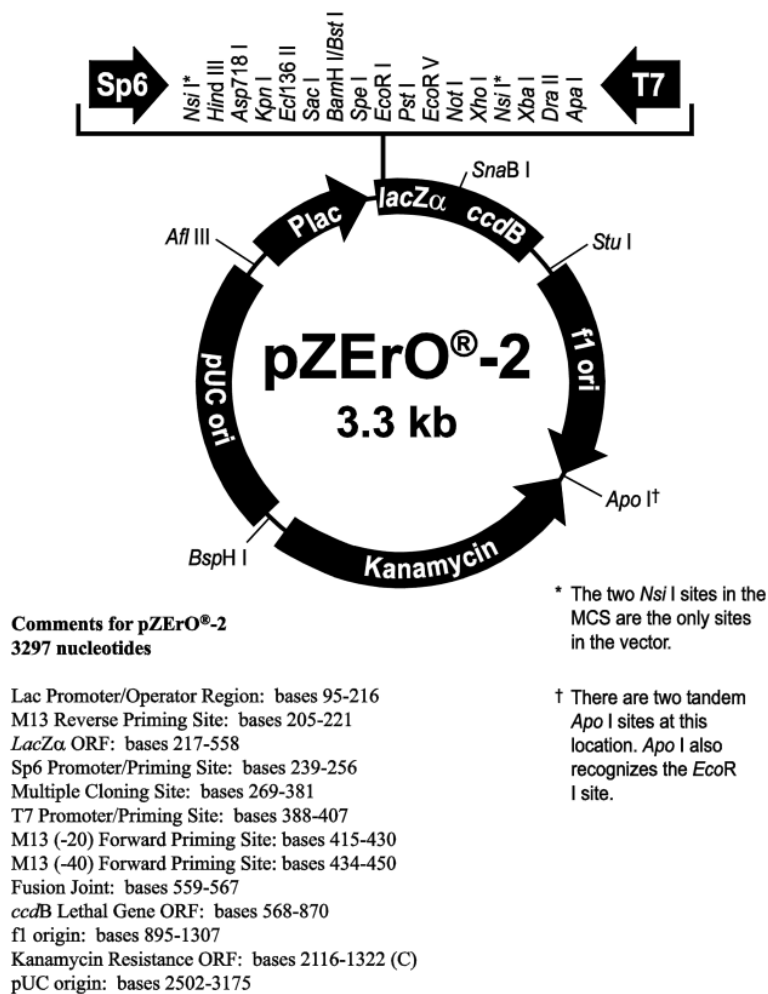


Fig. B.1 The Vector Map of pZero2.

```

631 atcaatgccaaatggaccagaacgaagccaagtaccgggtggaaaaagcccgaggagcagacaagtatgataagctgtgagctgcg
721 gctatcgctctacattccgggcttcggcctttgtggacctggtgctggcccaagatgcatcaatacctatttggccggacggtaaataaat
1 M
811 tattatagatgacaacaagtaccctgtcgaacattattccagggaagaaaggagccagaggtatgaaaagagaattcaaggatctgATG
2 R F E P Y E C L N S E I T G D Y D K A H A V K C V N G T F V
901 CGTTTTGAACCGTACGAATGCTTGAACAGCGAAATCACCGGAGATTACGACAAGGCCCATGCCGTGAAGTGCCTGAACGGCACCTTTGTG
32 G T E E H G V A S W L G I P F A K P P V G G R R F K A P E Y
991 GGAACCGAGGAGCACGGGTCGCGTCTGGCTGGGCATCCCTTTCGCGAAGCCGCCGTGGGCGGGCGCCGCTTTAAGCGCCGGAGTAC
62 V D A S D R V F K A R Y Y G K C A F G A Q A Y P D C V Q K L
1081 GTGGACGCCAGCGACCGGGTCTTCAAGGCCAGTACTACGGAAGTGCCTTTTCGGCGCACAAAGCCTATCCGGATTGTGTGCAGAAGCTG
92 S S E D C L Y L N I W A N T D D K T E K K P V M V W I H G G
1171 AGCTCCGAGGACTGCCCTTTACCTGAACATCTGGGCCAACACCGACGATAAGACCGAGAAGAAGCCGGTTATGGTCTGGATCCACGGCGGG
122 A Y V T G S G S Q I S Y S G A N L V Q K H S D I I M V N I N
1261 GCTTACGTACCGGCTCGGGCTCGCAGATCTCCTATAGCGGAGCCAACTGGTCCAGAAGCATAGCGATATCATCATGGTGAACATCAAC
152 Y R L N M Y G F M D F S S V P G G E N F K T A P C N G L L D
1351 TACCGCTGAACATGTACGGCTTCATGGACTTCTCGTCGGTACCCGGCGGCGAAAACTTTAAGACGGGCGCCTGACACGGCTTGCTCGAC
182 Q A M A L R W V H E N I A A F G G D P D N V T I F G Q S A G
1441 CAGGCTATGGCGCTCGGGTGGTGCACGAGAACATCGCGCTTTTCGGCGGCGACCCCGACAACGTGACCATCTTCGGGCAAAGCGCCGGC
212 G G S V S I L P V M K E A N R Y F Q K V I A Q S G S A T L A
1531 GGCGGCTCGGTCTCCATCTGCCGCTCATGAAGGAAGCGAACCGGTATTTCCAGAAGGTATCGCCCAGAGCGGCTCGGCCACCTGGCA
242 F P V D C E A A Q G K T K A L L E F T G C K T M D D I M A L
1621 TTCCCGGTGGACTGCGAAGCAGCCAGGGCAAACTAAGGCACTGTGGAGTTTACCGGCTGCAAGACCATTGACGACATCATGGCGCTG
272 S E E A L Q E A Y V E A V G K F T S C P Y Y G T E V L P E A
1711 AGCGAAGAGGCGCTTCAGGAGGCGTATGTGGAAGCAGTCGGCAAGTTACCTCCTGCCCTACTACGGAACGGAGGTTCTCCCGGAGGCG
302 P I E L Y R K G Y A K H I S I M A G T T A D E M R L F M G E
1801 CCCATCGAGCTGTACCGTAAAGGTTACGCAAAGCACATATCCATCATGGCAGGCACCACGGCCGACGAGATGAGGCTGTTTCATGGCGGAG
332 G P C L S L E E Q K L Y A R R A A G D A V P Y L K E E D K K
1891 GGCCCGTGTCTGAGCCTGGAAGAACAACAACTATACGCCCGGCGCGTGCAGGCGACGAGTTCCTTACCTGAAGGAAGAGACAAAAAG
362 Y Y E E F R R V C R D Q E P G L V E T E F I N E L M F R V P
1981 TATTACGAAGAATTCGGCGGGTATGCAGGGATCAGGAGCCGGGACTTGTGGAGACGGAGTTCATAATGAGCTCATGTTACGGGTCCCC
392 M L Q Q L D A Q S A F N K T F C Y Y W S Y P G S N P D M G A
2071 ATGCTCCAACAGCTCGACGCCCAAGCGGTTCAACAAGACGTTCTGCTACTACTGGTCGTACCCGGGCAGCAACCCGGATATGGGGGCA
422 A H S V E L L F V F D F R G V G T D S T F N G T N I P E E I
2161 GCGCATTCGCTGGAGCTCCTGTTCTGTTTTCGCGGCGTGGGACGGACAGCACCTTCAATGGCACAACATCCCGAAGAGATC
452 F T A V Q M W T N F A R C G N P S T D K V E W K A Y S V D
2251 TTCACGGCTGTTTCAGCAGATGTGGACCAATTTTCGACGCTGCGGCAACCCGTCACAGACAAGGTCGAATGGAAGCGTACTCGTGGAC
482 D Q N V L V I A G P D D I H I E Q G L L A D Q Y K A V L P L
2341 GACCAGAAGCTCCTGGTCATCGCCGGCCGGATGACATTATATCGAACAGGGCCTTCTGGCCGATCAGTACAAGGCCGTGCTCCCTTG
512 L G Y Y Q F M D K F F T P G Y L L D I V A A R Q Q N A *
2431 CTGGGCTACTACCAGTTCATGGACAAGTTCTTTACGCCCGGCTACCTTTTGGATATCGTCGCCGCGCCAGCAGAATGCCTAAgctctg
540 M N T K I K C N N G T F I G
2521 tagggaatataagacaaacttttgtaaaaaagagaagaagcaaatcacATGAACACTAAGATCAAATGCAACAACGGCACTTTCATCGGC
554 K E T D G L I I W K G I P Y A T Q P V G K L R W K K A L P A
2611 AAGGAGACAGACGGACTGATAATCTGGAAGGGCATCCCTACGCAACCCAGCCTGTGGGAAAGCTTCGCTGGAAGAAGGCAGCTTCTCGCC
584 A D D D G V Y D A T K P G H I P I G P V N D S M E T A E F G
2701 GCGGACGATGACGGCGTGTACGACGCCACAAAGCCCGGCCATCTCCGATCGGGCCCGTGAACGACTCCATGGAACAGCGGAGTTTCGGC
614 E D C L V L N I Y C N T G C T D S R K P V M V W I H G G G F
2791 GAGGACTGCCTCGTGTGAACATTTATTGCAATACCGGATGCACTGACAGCAGGAAGCCGGTTCATGGTATGGATCCACGGCGGGGGTTTC
644 C A E S Q A S P L Y D L T G I S R Q Y P D I L F V S I D Y R
2881 TGCGTGAATCCAGGCGTCGCCCCTTTACGACCTTACCGGCATCTCCAGACAGTACCCCGACATCTGTTCTGTTCCATCGACTACCGG
674 L G F L G F I N F E R V P G G K N F R E A G N L G L L D Q L
2971 CTCGATTCTGGGTTTATAAACTTCGAGCGGGTTCCGGGGGGAACAACTTCAGGGAGGCAGGCAACCTCGGCCTGCTCGACCAAGCTT
704 E A L R W V Q K N I A G F G G D P D N V T I F G E S A G S A
3061 GAGGCCCTCAGTGGGTACAGAAAAACATCGCCGGGTTTCGGCGGCGATCCGGACAACGTGACAATATTTGGTGAGTCCGCAGGCTCGGCA
734 S V T F L P L I N G S E G L F R K C I A Q S A N I A Y C D T

```

```

3151 AGCGTCACGTTCTCCCCCTCATCAACGGAAGCGAGGGGCTCTTCAGGAAATGCATCGCACAAAGCGCCAACATCGCCTACTGCGACACC
764 M E H G I H V T Q N F L T A A G C Q T M D E L M E L T T G E
3241 ATGGAGCAGCGCATCCACGTGACGAGAACTTCTGACCGCAGCGGGCTGCCAGACCATGGACGAACTGATGGAGCTCACCACCGGGGAG
794 L V D A Y Q K A S I I D K N C V L G T A N F P L L D G V T L
3331 CTTGTTGATGCCTATCAGAAAGCAAGTATCATTGACAAGAACTGCGTTCCTGGGAACTGCCAACTCCCGCTTCTGACGGCGTCACGCTG
824 P E D R A D M Y G M W G D E K R A K I D L M I G S N Q D E I
3421 CCCGAGGACCGGGCGACATGTACGGGATGTGGGAGACGAAAAGAGAGCCAAGATCGATTTGATGATTGGGTCTAACCAAGATGAAATC
854 R Y F V P F E G G E E G F A N T L R W I A K R D R A L L N E
3511 AGGTACTTTGTTCCCTTCGAGGGCGGCGAAGAAGGCTTCGGAATACGCTGAGGTGGATCGGAAGAGAGATCGCGCGTCTCAACGAG
884 Q E K A M Y D E F M A T L A N E S E G S R L E Q Y C N D I N
3601 CAGGAGAAGGCCATGTATGACGAGTTTCATGGCCACGCTGGCAACAGAGAGCGAGGGGAGCAGGTTGGAGCAATACTGCAACGACATCAAC
914 F R A G N T N M A I R H S A A G G N T Y M Y F M K K P V I T
3691 TTCCGCGCGGTAAACCAATATGGCCATCAGGCATTCTGCCCGCGGCGGAAACACCTACATGTACTTCATGAAGAAGCCGGTGATAACC
944 P Q L G A I H A A E I P Y L F D T C L E D P M G S G E I V G
3781 CCCAGCTTGGCGGATACACGCAGCCGAAATTCGTACCTGTTGCACTTGCCTGGAGGATCCGATGGGAAGCGAGAGATCGTTGGG
974 T D E A E G F R H V V K E M W V N F A R T G N P S T D K Y E W
3871 ACCAGCAGTCGAGTTCGCCACGTCTGTAAGGAGATGTTGGTCAACTTGGCCGAGCGGCAATCCTCCACCGACAAATACAGATGAGTGG
1004 K K F S G D D R Q T M V F D D A I G M Q K D L F G R E D L
3961 AAAAAGTTCTCGGGGACGACCGCCAGACGATGGTCTTCGATGATGCCATCGGCATGCAGAAGGATCTGTTGGAAGGCGCAGGATCTG
1034 M M S W A E R L G N G S S K R V C *
4051 ATGATGTCCTGGGAGAACGCTCGGGAACGGATCGTCCAAGAGGGTCTGCTAAcgtggcgtaagcgcaagtaacgattagaaattaat
4141 cttcaaatccaaagatttttgcggctaataatgttataataataagcaatgtttaaccattgcaatggtagtttcttt

```

B.1.2 N2 DNA Insert

```

1 gaggcacggctgtcggcgagttcgataccctgctgcgtggctgctgcccataacctcgggtgccgcctgcataccgcctggcgcat
91 tacctggcgctcctgatcgtcatggctgggttgatgttctgcgccccggcgctgatctacagccaggtctcccaggccctccagggtgac
181 gagcgccacctgctctaccagggttcttcaaggccttcatgggtgctcaccctgttcgccccggcgctggcctgggtgggtgctgacc
271 cgcgagagtcgccacgtggcgctggaggaatcgaccgcccagaccgccccgtgctgatgcaggaatccaggccaccgcaagaccgacgag
361 gccctgcagcgcgccaaggaagcctcgaggcgcccaacgcgcgaagagtcgctacgtcacgggtctgtcccacgagctgcgcacgccg
451 ctgaacagcatcctcggtatataccagatcctgcagcgcgacccggcgagcagcagtcgccagcaggacgccccgggcaccatcttcggc
541 agcggctcgcactcgtgtccttgatcgacggcctgctcgactggccaaagtcaggcgggcaagctcaacctggagcctcggagatc
631 ctttcccggaattcatcgagcagtcgacgagatgttcgccccaggccaggacaaggcctgagcttcgctcagcggcgggga
721 cgctcgccggcggtggtcgcggcgacgagaagcggttcgcccagatcctcatcaacctgctcggaacgcctgctgacacggcg
811 ggcggcgtggttactgggggtcacctacgcccgggaacgcaccttcgagatcagcgacagcggtcggggtacgccccggagcagctg
901 gagcggtgttccaaaccttcgagcgcggtgacctgctgcgcaggacaacggcctcgccctgggctgaccatcacccgcatgctggtg
991 accctcatggcgcgccctggtgctcagagtagcgaaggccaggcgccccgcttcctgctgctgctgttccttcggaggtacgcgtc
1081 cccaggcgctggtccacgtcgagcacaccatcacccgttaccaggggccgcgccccggcggtgctgctggtggatgaccacctgcaccac
1171 cgccgggtgctggcgccgcatgctcgaacccctgggtttcgaactggccaggccagcaatggccaggacgcatccgccagggtggcctg
1261 tggcagcccgacctgatcctcatggacctcgcatgcccgtgctggacggcctggagaccagcacctgatccgccgaatggcctctcc
1351 cgggcgcccatcatcgtcatttccgcaatgcttcgcccagcagccgaacgcagcgcgccccggcgctgagcagactacctggccaag
1441 ccggtacacacgcgcagctgctggagaagatcaagcgccacctggacctggaatggctggagcgccccgaggagcgccccaggcgcca
1531 ccgctgccccctgcaggcgccaagtccgccagtctggccgaactgcaggagctggcgccccctgggctacgtcaagggtatcctcgaatgc
1621 ctggagcgcacgcgagcgaggagccccgagcgcgccctacgtcgccacctgctcggggtggtcaagcgcttcagctcaacgacttc
1711 aaccggcgctcaaggacgccccgagcggcgccctgccccctggaaggagaaacccatgaatgcccctacagccccgctcccgccg
1801 gcgtcgtgctgatcgtcgatgacacccccgacaacctcgcgtgctcctcgatgcccctggagccaccggctacatggtgctggtggcca
1891 tggacggcgccagtgccctggaacgcatgcagcgcggcgctccgacgtggtgctgctgatcggtgatgccggcctggacggcttcg
1981 agacctgccccgggatcaaggccaggccgaactggccgacatccgggtgctgttcgatgaccgccccgagagagcagcagctggtgg
2071 aggccttcgccccggcgccgcatcgactatgtaccaagccccctcaagaccgacgaggtactggcggggtggcgccccacctgcgcaccg
2161 cccggaactccaggcagccaggagcctgcccggcgagccgccccaggccaggcgccgctggacctgcaccgctcagcagccgctacc
2251 agctgaccgctccccgaagaagtgcgctgacttgacggcattaatgaatgactatcgccaacgctgggggtattaaacatttcgtatt
2341 ggcaggttattcatttggcgcatgtacttccagctatttacaatcgctgcaagataaagataaaaacgatgtcagcagtatctgtt
2431 actggcttttcacgcaaaaggaagtttcgagataacacctggacggctggataaaagacagtaataacggtattgaaacgggatccgaaat
2521 ggccaaactcccgccataaggtgctctgcgtgtacggcgtgaaggaaaagaaaagcggtgtacggacccccactgtcgtgggtga
2611 agtcttgcaacttccggcaaacatcactttgatcgcgactacaactcactgacgacgaaactgctggatgcatcaagcaccggaagcgc

```

```

2701 gaccaccacgctttaagtgtgagtcgaagagaagggcgctctgcaacgggcgctcttctgaaccgtgagaagcaagggatgtggc
2791 caatagccaatatcctaataatgatctgtcctgtttgtgacgttaacgcttcactgttgatacatgacatggcgatatgctgtaactgct
1      M L N T A L R R P F A A C L L S L
2881 accttctcgttcagtaatacaaaaagaaggatttctcgATGCTCAACACTGCTCTGCGCCGCCCTTTGCGGCGCTGTCTCCTCTCGC
18   A C S S A M A A S S P Y S T M I V F G D S L A D A G Q F P D
2971 TCGCCTGCTCCTCGGCCATGGCGGCTTCTCTCCCTACTCGACGATGATCGTCTTCGGTGACAGCCTTGCCGACGCGGGCCAGTTCCCCG
48   A G G P A G S T L R F T N R V G P T Y L N G S G E A F G L N
3061 ATGCGGGCGGCCCTGCGGATCGACCTGCGTTTCACCAACCGTGTGCGGCCAACCTACCTGAACGGCAGTGGCGAAGCGTTCGGGTGA
78   S S T L L G G M L G V A P N D L Q A S T S P V R A A Q G L A
3151 ACTCATCGACGCTCCTGGGCGGCATGCTGGGCGTCGCCCCAACGACCTGCAGGCCTCGACGTCTCCCGTGC GCGCGCTCAGGCGCTGG
108  D G N N W A V G G D R T D Q I L A A I T T Q S Q V A N T T S
3241 CGGACGGCAACAATTGGGCGGTGGGCGGGGATCGAACGGACAGATCCTGGCGGCCATCAGACCCAGTCCCAGGTGCGCAACACCACCT
138  G V T T V F R T R P G Y L V E N N F R A D P N A L Y Y L T G
3331 CGGGGGTGACGACCGTGTTCGCGACCCGACCGGGCTATCTGGTTGAGAACAATTTTCGCGCGCACCCCAATGCCCTGTACTACCTGACGG
168  G G N D F L Q G R V I S P G Q A I S A A N N L A D G A Q V L
3421 GGGGCGCAATGACTTCCTGCAAGGGCGCGTAATCAGCCCCGCAAGCCATCAGCGCGGCAATAATCTCGCGGACGCGCGCAAGTGC
198  S Q A G A R Y I M V W L L P D I G Q T P A V S G T F L Q G P
3511 TGTCCCAGGCGCGCGCGCTACATCATGGTCTGGGTGCTGCCTGATATCGGTCAGACACCGCGCTCTCTGGCACCTTTCTGCAAGGGC
228  S T L L S G V F N Q Q L V N R L G Q I N A E V I P L N V P G
3601 CATCGACGCTGCTCAGCGCGTGTTCATCAGCAGTTGGTCAATCGCCTTGGCCAGATCAACGCAGAAGTGATCCCGTGAACGTGCCGG
258  L V R E V L D D P A R F G L A A D Q N L V G T C F S G N S C
3691 GACTTGTTCGGGAAGTTCTGGACGACCTGCCCGCTTTGGCCTCGCGGCGGATCAAAATCTGGTGGGCACCTGCTTCAGCGGCAACAGTT
288  T A N S V Y G I G G T S P D P S K L L F N D G V H P T V A G
3781 GCACCGCAACAGCGTTTACGGCATCGCGGCACCGCCCGATCCGAGCAAACTGCTGTTCAACGATGGCGTGCACCCGACGGTCGCCG
318  Q R L I A D Y G Y S I L S A P W E I T L L P E M A N G T L R
3871 GTCAGCGGTGATCGCCGACTACGGTTACTCGATTCTCTCGGCGCGTGGGAAATCACCTGCTCCCGGAAATGGCCAACGGCACCTTGC
348  A Q Q D E L R S Q W L A D W G N W Q N V G E W R A I V A G G
3961 GTGCGCAGCAGGACGAACTGCGCAGCCAGTGGCTGGCCGATTGGGGCAACTGGCAGAACGTTGGAGAATGGCGCGCCATCGTCGCAAGCG
378  G Q K L D F D A Q G N S A S G D G H G Y N L T V G G S Y R F
4051 GCGGTGAGAACTGGACTTCGATGCGCAAGGCAATTCGCGGAGCGGCGACGGCCATGGCTACAACCTGACCGTGGGCGGACGTACCGCT
408  A E D W R S G L V A G A Y R Q T L E A G P R D S D Y T L N S
4141 TTGCTGAAGACTGGCGCTCAGGCTGGTGGCCGGGGCTATCGGCAGACGCTGGAAGCCGGGCAAGGGATTTCGACTACACACTCAACA
438  Y I A T A F V Q F H T N H W W A D L A A S G G K L D Y D N L
4231 GCTACATCGCCACGGCTTTCGTGCAGTTTCATACGAATCACTGGTGGGCGGACCTGGCGGCATCAGGCGGCAAGCTGGACTACGACAACC
468  K R K F A L G V S E G A E K G D T D G N L W A L G R L G Y
4321 TCAAGCGCAAGTTCGCGCTCGGGGTGAGCGGCGCAGAGAAAGGCGACACCGACGCGCAACCTCTGGGCGCTGAGCGGACGGCTTGGCT
498  D L A E Q S S R W H V S P F I S A D Y S R V E V D G Y S E N
4411 ATGACCTGGCCGAGCAGAGCAGTCGCTGGCATGTCTCGCCCTTCATCAGTGCCGACTACTCCCGCTCGAGGTGATGGCTATTTCGGAGA
528  S T R A T A L N Y D D Q T R R S K R L G A G L Q G K F D V T
4501 ACAGCACTCGCGCACTGCGCTCAATTATGACGACCAAGCCCGGTCGAAGCGTCTGGTGCAGGCTGCAAGGCAAGTTTCGATGTCA
558  P Q T Q V F G E V A H E R E F D T D Q Q D V T I A L N S V P
4591 CGCCGACAGCAGGTGTTTGGCGAAGTGGCCACGAACGCGAATTCGACACCGATCAGCAGGACGTGACCATTCCTGTAACAGCGTGC
588  G V D F N L Q G Y E P Q R S L N R A S V G L S Q K L T Q D L
4681 CGGGTGTGATTTCATCTGCAGGGGTATGAGCCGCAACGCAAGTTTGAATCGGGCAGTGTGGGATTGAGCCAGAAGCTGACGCAAGACC
618  T L R A G Y N W R K N D D V T Q Q G V N V A V S L D F *
4771 TGACGTTGAGGGCCGTTACAACCTGGCGCAAGAATGATGACGTGACGCAAGGGGTGAATGTGGCAGTCAGTCTGGATTCTAAcgac
4861 catcagaccgcggtcgcgggcaagcgcgctcctacagaagggtgggggtcaggcttggtgctcttgttccgacgcgccaccgacctgaa
4951 catcgcgcggcggtttgttcagggtttcttccattccagcgccggtactgagtcggcgacgatgccgccaccgacctgcacgtgcagctc
5041 gccgttcttgatcacgcgtgtacgaatggcaatcgcggtgtccatgttaccgttccagccagataaccaccgcgccgctagacgcc
5131 acgtttgaccggtcca

```

B.1.3 N7 DNA Insert

```

1  agaacaggaaaaacggcaccaggcgggcaacaagtggatcggcacggcgggcacttcgccattcggacattcgggctacaaccgggaagg
91  tgtccgtatcgcgcgagggaaagcatggcgggccatcaaagtctgggaaaagcgaggtttgccaatctcgacaacacacgtgaact

```

```

181 gggcaccgcgaacatcaaggtcgcgctccgcccgtcgggcgtttcgcgctgaggagcgggcgaggaaactggatcttgacgccagat
271 ccatggcaccgccaaagcagggtcgctcgacatccatatcgcgccagagcgccacaacgcggtcaaagtactactgttcccttgacgtcgg
361 cggttcgatggaccgcgcatatcaaactgggtcgaagaactgttcagcgcgccagcgaggttcaagaacctgaatttttttacttcca
451 caactgcctttatgaaggtgtctggaaggacaacccgcccgttttcggagcagacgccaacgtgggacgtcctccacaagttcagcca
541 cgactacaaggtgatctttgtcggggatcgggcaatgagtcctatgagatcagccatcccgcggttcggtagagcatttcaacgagga
631 gtcgggcccggcctggatgcagcgcggtggaacacatcccgctacgggtcgtggaatccacgcccggaaagacaatgggaatattc
721 ggcttcgactcggtgatccaccagcttgtcaacggatcaatgtatccgctgacgctggaggggctggacgatgaatcggggaactgac
1 M G D A I E V T G L Q V P A R T I P T P
811 acgtaagaacattaatctgagggaatgaATGGGGGACGCGATTGAAGTGACAGGGCTGCAAGTGCCGGCCCGGACCATCCCGACGCC
21 N T I S S E A Q A F L S R G L P I M P P E I P H T D K D R W
901 CAACACGATAAGTTCCGAAGCGCAGGCTTCTGTCTCGTGGCTCCCATATAATGCCGCCGAGATCCCGCACACAGACAAGGATCGTTG
51 R G Y V A Q V E A Q I V R V A E M R A R A F P A A I S E H R
991 GCGTGGCTATGTGCGCAGGTGCAAGCGCAGATCGTCCGGTGGCAGAAATGCGTGCTCGCGCCTTCTCGCGCCATATCGGAACATCG
81 L G S T T L Y E V T P D T L D P A D E D K A I L H I H G G A
1081 CCTGGGCAGCACGACACTCTACGAGGTGACCCCGACACGCTCGATCCAGCCGATGAGGATAAGGCGATCCTCCATATTCATGGCGCGC
111 F G I V G G K S A A Y T A G T I S S L A G I R A F S P D Y R
1171 GTTTCGTGCGGGGGCGGAAAATCGCGCGCATACGCGCCAGTTCAGCTTCCAGCCTTCCGGGATACGCGCTTCTCTCCAGACTACCG
141 M P P D H P Y P A G L D D C V E A Y R F L L E R Y D P S R I
1261 GATGCCGCCGATCATCCTTATCCGCCCGGACTGGACGATTGCGTGGAAGCCTATCGCTTCTTCTGGAACGCTATGATCCGTCGGAAT
171 A L E G S S A G A N L V A A T I L R A R D E G L P L P G A C
1351 CGCGCTGGAAGGATCATCGGCCGGGGCAACCTTGTGCTGCCACGATTCTGCGCGCAGGGACGAAGGATTGCCACTTCCCGGTGCATG
201 S L H T A G V D L T H S G D T F A T N E V I D I V L R G P Q
1441 TTCGCTCCATACCGCGGTGTGGACTTGACCCACTCTGGCGACACCTTCGCAACCAACGAGGTGATCGACATCGTCTTCCGGGGCCGCA
231 P E T M L L Y A G G H D M R D P Y L S P V F G D V T K G F P
1531 GCCGGAACCATGCTCCTGTATGCAGGTGGCCATGACATGCGAGATCCCTATCTTTCGCCAGTTTTCGGGGATGTGACCAAGGTTTTC
261 P T I L V S G T R D L L L S P T V L M H R A L R R A G I E A
1621 TCCACGATTCTGCTATCAGGCACGCGGACCTGCTTCTCTCGCCACGCTGCTGATGATCGCGCCTCGCGCGCAGGGATCGAGGC
291 D L H V F E A M P H G G L G G A S P E D R E L Q L E I A S F
1711 TGACCTGCATGCTTTGAAGCGATGCCGCACGGCGGCTTGGCGGCGCTTCGCCTGAAGATCGTGAAGTGCAGCTGGAGATCGCAAGCTT
321 I R R H L T R T C A *
1801 CATCCGCGTCATTAACAGAACCTGTGATGActgatcgacacgtcggttagtcgcttgcgggatcgagcgaaacctaggactac
1891 gcagccgaggcggtgcctttggcgaccgattccatcggaacagggcctccctgaatggatagcaacttctcatctgcctgctg
1981 accgctggcataaacctgatcggcacgtcgcttatcgcccggtatcgccggtgtggaaccagggcgattgccatgagctttgcctg
2071 ttcaacgtgctggtgctcgttttcgcgacgtccaacgctttcttggaccgttccttgccaaacgtattgaaacgcggtttcggaaggc
2161 gggggagaggcattgtcggcgattttcgaatggtgctgtgtgagccacgattgagtggtggttgggcatcgtgctggtcccaactggc
2251 cagcggtggttgcgcgctatcagttatttccagaccacgctccaccaccaagatgttgctgcgaagcgcgactccaaggggcgctg
2341 cggaaccttcgcatccacacgccccacgttagtcagatcaaggaaactggccaagccacgcggtggtggtggtggttattgctg
2431 gccaactgccttgctcaagcgcttattacagtggtgctgctgcttgcgtttatgcgggctatctcaatccggaataccgggtgacggca
2521 tcgagcttttcggcggtgatcaacggcttcgccaccatcgtcgtttttgcctttatcgaccacaactctcggtaatgaccgacgacgta
2611 ggggaggcggggttagcgaacctgttcgccggagatcgtgtggtatttccttcagccgtctcgtcggaactatgcttgcgaggcg
2701 ctggtcgtccgctccgcaatggtaatcgctggatcgccaactacgtatagagcaggacgcctccgatttcgtcgccccaacggccttcg
2791 tccggttaaggtaagcctatcgagatcaccggcgaggcgggcggtcatccaccattcggaagagcgggcgccggtgatgacgtcg
2881 atcgcaagccgatcgaactcgtattcatagggaataatcgccagatcgctagtcgggtagagcctgcgaatgctttgaaggtgagc
2971 gcctgaagccgccacggcggaacgcggcatgatcgtaaaaatgtccctccgcatggatcatcaagcccgaacacaggctcttggcatgt
3061 ttgtgaaaggtcggttcgaaccagcattcgcggttgcgcatcccgccagcattgcggcgaaacgattgcatgtcggtcagcacggcc
3151 tttgatcaacatcaggagcaccgaacagcactccagcgcggggttgcctccctaccctcgctcaaggtcgtgacctccagcatgacc
3241 gtgcctgataggcggggcatgttaaggttggcggttgggacttggattgatccagatcggttttctgcccacaaagcccggt
3331 gttgcaccgggctgccaatcagccatctcaatgaccggtaatcgacatcgagcccggaagtgatcgacgaaccagcgccccatctcacc
3421 agtgtcagtcctatcgccatcggcctcgcccgccgacaaagctttcctgtccggggatcagcagctccctcgatcgggcggtccg
3511 gcaggattcgccctgtccaggaaccatactgatttaccgcgtcccgccgagcttcgagcacgtagagcaacgtcgtgacgaagtgtag
3601 atacggcagcaagatcctgaaggtcgaacaggaacacatcgggggtgtccatcattgccgttgcggcgggcgaaacctcgccataaagg
3691 ctgaagatgggaatgcccgtagtgcggatcgacctcatccgcgtctcgacctgttgcctgcttgcgccccagccgtgctcggg
3781 ccaatgcccagtcaggttgatgtccgggcatgacgccagcgcatcgagcgagtgctcaggtcttcagtgaccgacgctggatgagcg
3871 atcagggcccaccgcttgccaacaagggcttgcgagttcagggtccgacgagcgatctatgcccgaatttcagttcgtcttccgggtg
3961 caggaagatggagaacgaagcaatcgcatgatggaagtcgggcttaccgcggtatgggcaagcgccagtagcttttgggtccggtcgg

```


4051 tctcctcaatgacggcagacagggcaatgcagcgctcaccgagcagcggtatcgtggct

B.1.4 N11 DNA Insert

```

1  tattaacactaaaccgcaagttctgaagaacctgctgggtttttaatttcctcgttcttttgccaaaggttatacaaaacttgagtat
91  ttttaatatcctcggccattttcccgctcactgttgaaatattgtttttatatattataatttcactcactacaataagttgagctta
181 attctaagctcatttttccagaaatccttaaaatgaatacatgcttatagcaatggaagttttttgcatgcttttgctctgaaatat
271 cttttttattacgggctaaacgaaagaaaaagcaagaagcccgtaaaagcaagagacaccagccgagaaactacgggctaaacaagaaa
361 aagcaaaaaagcccgtaaaagcaagtgccagccgagaaactacgggctaaacaagaaaaagcaaaaaagcccgtaaaagcaagc
451 ggcaccagccgagaaattacgggctaaacaagaaaaagcaaaaaagcccgtaaaagcaagaagcactggccgagaaactacgggctaa
541 aacaagaaaaagcaaaaaagcccgtaaaagcaagaagcactggccgagaaactacgggctaaacaacaaaaagaaaaagcccgta
631 aaaacaagaaactatcaaaattataagagaaaagggggttaactggaaaaactacagagaaaaactcatacttcaactaaatgaataa
721 atgagttacttgcataatcaaatagaaatctttcaaaattacagaatattccgaagcgtctggttaaatcgagtaaaagcggaggtatgcg
811 ttcagtactattggtttgataagacgacaaataaaaagatatatacatcactaaagatcataaagaagaattgagaaaaacccttcaacgtg
901 actacgaaatttcctgcaaaaaaactgaacgccctcaaaagaaaactatcgaaattcctgaaaacatatgatatcgacgaaatagaaa
991 aggttttattccaaccttcggaagcaagaaaaatacttgaactccaatagtgatacaaaaggaagatttgaagaagtgaggaaagcgg
1081 tagagtatgagccgatggaatcagcgataatattgaatttgatcatggttaacggagtttaaggtcagatcaaatccgagttgattattg
1171 ctaatatgctggagcaaaacggagctgcttatagatacgaatatccgcttatgcttaacggattaggaacagtcagaccggattttctct
1261 gcctgaataaacgaaccggaaggaatatgtttgggagcattttggcatgatggaataatagcgtatgccataaaaaacatagctaaaa
1351 tccaaacctatgagcagaacgggttctcgcgggaaaaaacatgatcatgacatttgagtcacgatgactccgctcagttcagccacaa
1441 taaaacaaatgatagaagaatatattgttataaaattgtattaccattgggtgacagaatttggttggtctaatcaatggatgatattgtaa
1      M A N I S V R F Y S N C L R R
1531 aataattaagaaagtattccaaccaccacagaaaggaagaatgacATGGCAAATATTCTGTTTCGCTTTTATTCAAAGTGTGTTGAGAAG
16  F T T F N M Y L P N D I R E E P N E S E Y A N R P I K T L
1621 ATTTACGACATTTAATATGTATCTGCCAATGATATAAGGGAAGAGGAGCCCAACGAGTCGGAGTACGCGAATCGTCGGATTAAGACGCT
46  F L L H G Y T G D A D N W V P W Y L A D K Y N F A V V I P N
1711 GTTCTGCTTCACGGATATACGGGCGACGACAGATAACTGGGTGCCGTGTTATCTGGCGGATAAGTATAATTTGCGGTTGTGATTCCAAA
76  G E N A F W L D G I S T G H A F C K F V G E E L M D Y V R R
1801 CGGAGAGAACGCATTCTGGCTTGACGGCATTTTCAGCGGGTCATGCTTCTGCAAATTCGTTGGGCGAGGAACCTCATGGATTATGTCAGGAG
106  V F G L A K T K E D T Y I M G L S M G G F G A L H T A L Y Y
1891 AGTTTTCGCGCTCGCAAAACAAAGGAAGACACCTACATCATGGGATTGTCCATGGGCGGATTCTGGCGCACTGCACACCGCTTTGTACTA
136  P D K F G T A T A L S P A L I V H E V A S L K E G E G N G I
1981 TCCCGATAAATTGGCACCGCGACAGCACTTTCTCCCGGTTAATCGTACATGAAGTAGCGTCCCTAAAGAGGGCGAAGGAAACGGGGAT
166  A N Y E Y F R E C F G D L T K V L E S R N N P E T L I K E I
2071 AGCAAATTATGAGTACTTCAGAGAATGCTTCGGAGATCTCACGAAAGTGCTTGAAAGCAGGAACAATCCCGAAACGCTCATAAAAGAAAT
196  K A E G K D C P K I F M A C G T D D F L I E N N R D F H K F
2161 AAAAGCCGAAGGCAAGACTGCCCTAAATATTCATGGCTTGTTGGGACAGATGATTTCCTGATCGAGAATAACCGAGACTTCCACAATTT
226  L E S E G I E H V Y A E E K G D H N M E F W D K Y L R I F I
2251 CCTTGATCCGAGGGAATTGAACATGTTTATGCGGAAGAGAAGGGCGATCATAACATGGAATTTTGGGATAAATATTTGAGGATATTTAT
256  P K M F E *
2341 ACCTAAGATGTTCAATAAActcagggtgctttggctaccgggtgacactgggggacggggttttaattgtaatagaaaaatcca

```

B.1.5 N13 DNA Insert

```

1  gttaccctctggcgccgggcatcaattccaattccttgatgtgttacacaaaaaatcagatcagatgaccgaccagaggcttattaatt
1      M
91  gttgctgttgctgttaatatgactcggtcgggtttgaacctaccctcaggctcatcatggataacaacatataaagagataaggatAT
2  K K K L I Y A A V V S A L L S G C G G S D E N K G D T S S Y
181 GAAAAAGAAGCTAATTTACGGCGCAGTCGTGAGTGCCTGCTGAGCGGTTGTGGTGGCAGTGACGAAAAACAAAGCGATACCTCAAGCTA
32  L D Y L L S G T N A A R P S A L A A R A S D G T L K F S T E
271 TCTGATTATCTGCTGAGCGGCACCAATGCGGCGGCCCGAGCCCTGGCCGCCCGTCCAGTGACGGCACCCCTGAAGTTCTCCACCGA
62  T A D L S N P V S A L S T L D G W S T T Q A I Q I V P V T A
361 AACGCGCGATCTCTCAATCCAGTCTCAGCCCTCTCCACCTGGATGGCTGGTCCACTACCCAGGCGATCCAGATAGTCCAGTGAAGTGC
92  S G I T V K A P S A A E F G A S V A P L Y L L E L E F D S A
451 CTCGGGCATCACAGTCAAGGCGCGCTCCGCGCGCAATTCCGTCCTCAGTGGCTCCGCTCTACCTGCTGGAGTGAATTCGACAGCGC

```

122 A L R P S G V K K V L A Y G V D F V V A E S A G K L N L V P
 541 CGCCCTGCGCCGAGTGGTGTGAAGAAGGTGCTGGCCTACGGGGTGGACTTCGTGGTCGCCGAGTCGGCGGGCAAGCTGAACCTGGTGCC
 152 L K P L N P S S Y Y M I V A T D S L K D S S G N P L R A G S
 631 GCTCAAGCCGCTCAATCCCTCTTCTACTACATGATAGTGGCGACCGATTGCTCAAGGACAGCAGCGGCAACCCCTGCGGGCGGGCAG
 182 D Y S N Y K S T T G S N A Q E Q T I S G L I A L Q E G L F K
 721 CGACTATAGCAACTACAAGAGCACCACCGGCAGCAATGCCAGGAGCAGACCATCAGCGGCCTGATAGCGCTGCAGGAAGGGTTGTTCAA
 212 A A T G I T S D H V I F S D W F G T Q S G A D V L V A V K G
 811 GCGGCCACCGGCATCACCAGCGACCATGTGATCTTCTCCGACTGGTTCGGTACCCAGTCCGGCGCCGATGTGCTGGTGGCGGTGAAGGG
 242 A A A S V L K S P T L D A A A L W K Q D A L G N T S L P G T
 901 CGCGGCCGCTTCCGTGTCTAAGAGCCCGACCTGGACGCGAGCCGCTGTGGAAGCAGGATGCCCTGGGCAACACCAGCCTGCCCGGCAC
 272 Y T L A V T G S N T F L T Q L D A E Q F L P Q E Q K D A I A
 991 CTACACCTTGCCGTGACCGGTAGCAATACCTTCTGACCCAGCTGGATGCGGAGCAGTTTCTGCCGCAAGAGCAGAAAGATGCCATTGC
 302 A A V E V N P Q L K G L A G M T Q V F T G T V K L P Y F L S
 1081 CGCCGCCGTTGAAGTTAACC CGCAATTGAAAGGTCTTGGCGGTATGACCCAGGTCTTTACCGGTACGGTCAAGCTGCCCTACTTCTCTCTC
 332 S P A T A G S W D K A R T Q S W H G A I P S L Y A I A N A L
 1171 CTCTCCGGCCACTGCGCGCTCCTGGGACAAGGCCAGGACCACTGCTGGCATGGTGCCATCCCAGTCTCTATGCCATGCCAATGCGCT
 362 K A Q D A E V I G G L V G A G V D P A L L G E L I A D P S R
 1261 GAAGGCACAGGATGCCGAGGTGATCGGCGGTCTGGTGGGCGCGGTGTGGATCCGGCCCTGTGGGCGAGCTATCGCCGATCCGAGGCCG
 392 Q A E L L A E A S K L I G V T L T S G G K A L D P E Q N I G
 1351 TCAGGCCGAGTGCTGGCGGAGGCGAGCAAGCTGATCGGGGTGACCTCACCTCCGGCGGCAAGGCGCTGGATCCCAGAGCAGAATAGG
 422 R F N P L P K L E E V Q S V P M R I F A A T N D L K T I T D
 1441 CCGCTTCAACCCGCTGCCCAAGCTGGAAGAGGTGACGTCCTGCTATGCGGATCTTCGCCGCAACCAATGACCTGAAGACCATCACAGA
 452 V I I Y Q H G V T S V K E N A Y A L A L G Q I G A G A Q A S
 1531 TGTGATCATCTATCAGCACGGCGTCACCTCGGTGAAGGAGACGCTACGCCCTGGCACTGGGTGAGATCGGTGCCGGTGGCCAGGCGAG
 482 K N V A V V V I D H P L H G E R G F A L T G S P D S V T T D
 1621 CAAGAAGCTGGCCGTGGTGGTATCGATCATCCGTTGCACGGGAGCGTGGCTTCGCCCTGACCGGCAGTCCCTGACTCGGTCAACACGGA
 512 K N P T P Y L N V S Y L T V A R D N L K Q S V A D L L G L R
 1711 TAAGAACCCGACCCCTACCTGAACGTGAGCTACCTGACGGTGGCGCGGACAACTGAAACAGAGCGTGGCGGATCTGCTGGGCGCTGCG
 542 L A V G L A N A K G A I G Q K G L K V H F L G H S L G A I A
 1801 TCTGGCGTGGGTCTGGCCAATGCCAAGGTGCTATCGGTCAAGAGGGCTCAAGGTGCACTTCTGGGTCACTCCCTGGGTGCCATCGC
 572 G A N L L A V A N Q P I G N A Q A D A L F K F T T G G L A M
 1891 CGGGGCCAACCTGCTGGCGGTGGCCAACAGCCCATCGGCAACGCCAGGCGGATGCCCTGTTCAAGTTCAACACCGGCGGCTGGCCAT
 602 P G G G I A P L L L N S P T F G P T I Q M S V L T G G S A A
 1981 GCCGGTGGCGGATAGCCCGCTGCTGCTGAACTCGCGACCTTCGGCCCGACCATCCAGATGAGCGTGCTGACCGGTGGCAGTGTCTG
 632 L K T A T A Y A P N C K T A P A P T C F V N E F L P S L D A
 2071 CCTGAAGACGCTTTCACCGCCTATGCGCCGAACCTGCAAGAGCGGCGCGGACCTGCTTCGTCAACGAGTTCCTGCCGAGCCTGGACGC
 662 A T Q A S A A G T L Q S Y S F A A Q S V L D S A D P I N L G
 2161 CGCCACCCAGGCGAGCGCGGCGGTACCTGCAGAGCTACAGCTTCGCGGCCAGTGGTGCTGGATTGCGCGGATCCGATCAACCTGGG
 692 R G I A A D F P L F A T E V V G D G A L S L S D R V I P N S
 2251 TCGCGGCATAGCGCGGACTTCCCGCTGTTTGCCACCGAAGTCGTCGGCGACGGCGCCCTGAGCCTGTGCGATCGGGTTATCCCGAACAG
 722 I A T A P L G G T E P L F K V L A L Q P L S A T G A A N H H
 2341 CATTGCCACCGCCCCCTGGGTGGCACCGAGCCGTGTTCAAGGTGCTGGCCCTGCAACCCCTGAGCGGACCGGTGGCGCAACACCA
 752 A T R F V A G G H S S L L A P D E N F D P T G A V T T E M Q
 2431 CGCCACCGCTTCGTGGCGGTGGTCACAGCTCGCTGCTGGCACCGGACGAGAACTTCGATCCGACCGGTGCCGTCAACACAGAGATGCA
 782 T Q F G S F F A S G G T A V K V T D A S L L K Q *
 2521 GACCCAGTTCCGCGAGCTTCTTCGCAAGCGGCGGCACCGCGGTCAAGGTGACCGACGCCAGCCTGCTCAAGCAGTAAgacctgtcgcccaat
 2611 aaaaacgcccgttggggcggtttttgtttttggggtaataagcggtgggctgactgtagcaaacgcctcacgctggtctgtgatgc
 2701 ctgtccagccagccttgtagcaggccgagaccagcagaccagcagatgggcatgttgcggtcgctctgggtgcccagcagattgcag
 2791 ccgctcaatgactacaacaactcaggcgggctgtgatgcctgtccagccagccttgtagcaggccgaccagcaggccgaccagatgcg
 2881 ccatgttggcggtgggagtgccgagcatgtcgaagaagccgagcaccagcagatcagcatgaagcccatcaggcgccggcgcatgtgga
 2971 ggccgagggcggtgtagcctggcccgcatccagctgtagccgagcaggcatagaccacgcccggagaggccgcaaacgcgggccac
 3061 ttacgaagaactcggcgatattgggtaattgtgcgcaacgatgagcaggataaacagcttgccgctgcccagcggctgctgcg

B.1.6 N16 DNA Insert

```

1  gccctgatcatccatatgtttgtacaggtcttcccgttggtcaacatgggtgttgcgatcctggcgctttccatgtacggcattatcctg
91  tcggatcagatcgaccaatatctgctcagcagcgagagatcgcccatcagcacgccaacatcctgggtgcttcagatgaggccgcacttc
181  atccacaacacccctgatgaacatctattacctctgccggcaggacctgctgaagcccagagggtcacccctgaatttcaacacttttctt
271  gaaaggaatctgaacgctctggccagcaatgaaatgatcccccttttcagaggaactggaacatacacgggcctatctcaatgtggagctg
361  gttctgcatgacgaaaaccttcttctgattatgacatgacctatatccgttttcgaattcccccgctgactttgcagcctatagtggag
451  aatgccatcaagcatggcatgaatccggatgccgccccgctccacatttccataaaaaccagaaagacggattccggcagcgagatcatc
541  gtttcagacaacgggcccggggtttgaacctgcatctgaaggcgggccccacattgctgttatgaatatccggcagcggtggagatgatg
631  tgcggcgggaaaatggcgatcctgccccggaaggcgggcggtgtgtgaaactgacgattccctgatctgtcgcagtaaacgggact
721  gcgttacagaccgcctgaaacgcgagagggttttagcgaaaaaacaggatcaggcatgactgcttccccttatgttgacattttggaa
811  cactccttgtcccggcgggaacagatttataaaattatagtgaacgcataaaaacttggaacttcattgtaaagcaaaaaacatataagg
1  M Q F R A N S K I T D G N Y D R S L A V K C I N G T F
901  aggaacgaagaATGCAGTTCAGAGCAAACAGCAAGATTACCGATGGCAATTATGACAGGTCTCTGGCTGTCAAATGCATCAACGGCACTT
28  V G R K E D G V I V Y R G I P F V G K Q P V G E N R W K A P
991  TTGTGGCAGGAAAGAACGGGTATCGTTTACAGGGGATTCCCTTCGTGGGAAAGCAGCCTGTGGGGAGAACCGTTGGAAGGCAC
58  V D V V P D D G V Y E A Y Y K G K S P C Q H K D F S D V E D
1081  CGGTGGATGTTGTTCCGGATGACGGCGTATACGAGGCGTATTACAAAGGGAAGAGCCCTGCCAGCATAAGGATTTCAGCGATGTGGAAG
88  T L I N Q G E D C L Y L N V W K A D D D S T A K K P V M V W
1171  ATACCTTATCAACAGGGAGAAGATTGCTTTTATCTGAACGCTTGAAAGCGGACGATGACAGCACAGCGAAGAAGCCGGTCATGGTGT
118  I H G G A F E F G A A A F S L F E C D N F L R E N P D I I I
1261  GGATCCACGGCGGCGCCTTTGAGTTTCGGCGCGGACGCTTTTCACTGTTTGTGAGTGCAGATAATTCCTGAGAGAGAATCCGATATCATT
148  V T V A Y R L G I F G Y F H L S H L P D G G D F P D A Q N L
1351  TCGTTACTGTGCGTACCGGCTGGGTATCTTTGGTTACTTCCACCTGTCCACCTGCCGACGGAGGGGATTCCCGGATGCGCAGAACC
178  G P L D Q L M G L K W V H E N I A G F G G D P D N V T I Y G
1441  TGGCCCTCTGAGTACGCTTATGGGCTTAAAGTGGGTCCATGAAAACATTGACAGGCTTTGGCGGCGATCCGACAAATGTGACGATCATG
208  E S A G A G S V S L L P L L K G S H A Y F K R V I A Q S G S
1531  GAGAATCTGCCGCGCGGAAGCGTATCCCTGCTTCCGCTTCTCAAGGGTTCCACGCGCTACTTTAAACGGGTTATTGCCAGAGCGGTT
238  P T L T R S P E E A I D C T K V M M E V L G C K T A A D L M
1621  CCCCTACCTGACAAGATCCCCGAAGAGGCTATCGATTGCACGAAGGTAATGATGGAGGTCCTCGGCTGCAAACTGCTGCCGACCTGA
268  K V D A R T L A D E S E A V R L R I C P E R D G R W L P T D
1711  TGAAGGTAGATGCCCGACTCTCGCGGACGAATCCGAAGCGGTAAGGCTTCGCATATGTCCGGAACGGGACGGAAGATGGTCCGCGACGG
298  P Y E A Y A G G A A K D I D L M F G C N K N E F D W F A A A
1801  ATCCTTATGAGGCTTATGCCGGGGAGCGGCGAAGGATATCGACCTTATGTTTCGGCTGCAACAAGAATGAATTCGACTGGTTTCGGCGGTG
328  M G E E G I K M L A A D R K E R K L D K L P E K E K A L L E
1891  CGATGGCGGAGGAAGGCATCAAAATGCTGGCTGCCGACCGGAAGGAGAGGAAGCTCGACAAGCTGCCGGAAGGAGAAGGCCCTGCTCG
358  S F C K D V K G E G C E G E C R L F D Q L W F N A P V I R I
1981  AGAGCTTCTGCAAAAGATGTGAAGGGTGAAGGCTGCGAGGGCGAGTGCCGCTGTTCGACCGCTCTGGTTTAAATGCGCCGGTCATCCGGA
388  S E S Q A A A G G N I H T Y F F T A E P G H G V E L E I I F
2071  TCTCGGAGAGCCAGGCGGCGGCGGCAATATCCACCTATTTCTTACGGCGGAACCCGACACGGTGTGGAGTGGAGATCATTT
418  D H Q N T D R E L G R V F D E T F G K T V R K M W V Q F A K
2161  TTGATCATCAGAATACGGATCGTGAACCTTGAAGAGTTTTTGATGAGACCTTCGGAAGACTGTGCGGAAGATGTGGTCCAGTTTGCTA
448  T G D P S L S A D I S P D G K A K K W P A Y D P E T R Q V M
2251  AGACCGCGATCCCTCCCTCAGCGCGGACATTTCTCCGACGGAAGGCAAGAAAGTGGCCGGCCTATGACCCGGAGACCCGCGCAGGTGA
478  V L D E F D I H S E K E S A V K I V D W D R T Y D L T K Y Y
2341  TGGTCTGGATGAGTTCGATATCCATAGTGAGAAAGAATCGGCGGTAAAGATTGTAGATTGGGACAGGACCTATGACCTGACCAAAATATT
508  L F *
2431  ATCTGTTCTGAaccttgcggcccgcgcccatgacaggattaggggcgccggcgcatcaatatgaagcaccaatgtatggcttgagaatc
2521  tacaaggaggaaaatcagcatgaaggcagaccgattttcaaaaacgcgaaaatcttcacagctgacagggaacccgcaggccactg
2611  cccttgactgaagaacggcaaatatttttggcgacgaggaggacttaaggactttgaggcgagggtcaccgacctcgcgggga
2701  agttcattatgccggtattattgacagccatgtccatgtgacgacaggtatcggtttgcatacatggacagggagaatatattgtat
2791  gctccagcaaaaaggaagccctggactttatggcgagttgattaaaagcaatccgggaaggagcgctacaggtttgttctggagcgaa
2881  aattcctgaagggggaaattctcacaaggaagatttgattt

```

B.1.7 N18 DNA Insert

```

1 gcctgccgtcccggctgccttgcaacccctctgcccaggcctctcagcgctgcaacgaacgtctcgtactttttgttcacaaccagat
1 M T F S L
91 acaacgtcataaaaaccgtctagcttcagacgacgcgccccaacgctgctccgataacaacaatcaaggtgcaggcATGACATTTCTCT
6 R Q R L P L A I A L A T A L A A T V Q A A P N P Y S G F T V
181 CAGGCAACGCCTGCCGCTGGCCATCGCCCTGGCGACGGCGCTCGCGGCAACCGTGCAGGCGGCTCCCAATCCCTACAGTGGCTTCACCGT
36 F G D S L L D A G Q F P D T G V T G A S L R F T N R V G P G
271 GTTCGGCGACAGCCTGCTTGATGCCGGGCAATTTCCCGATACCGGGGTGACCGGCGCCAGCCTCAGATTACCAACCGGGTCGGCCCCGG
66 Y S A A G G A V T G P V S S I L L G Q Q L G F D A R T L D A
361 CTACAGCGCCGCCGGCGCGGTGACCGGACCGGTGTGTCGATCCTGCTCGGCCAGCAACTGGGTTTCGACGCCCGGACCCCTGGACGC
96 S T S V I N R L L G L A A G D N W A T G G Y T T A Q I R D S
451 CTCGACCTCGGTGATCAATCGGCTGCTGGGCTCGCCGCGCGGCAACTGGGCCACCGGTGGCTACACCACGGCGCAGATTTCGCGATTTC
126 I T A A N G S V V A A N G L T L R S R D G Y L P G L A S Q G
541 CATCACCGCCGCCAACGGCTCGGTGGTGGCGGCAATGGCCTGACCTCGCGACGGCGACGGCTACCTGCCCGGACTCGCCAGCCAGGG
156 L R L D P N T L F Y I S G G G N D F L Q G L V T S P A S A A
631 GCTGCGCTGGATCCCAACACCCTGTTCTATATCAGCGGGGCGGTAAACGACTTTCTCCAGGCGCTGGTCACTCGCGCGCAGCGCCGC
186 A A A N R L G D G V A A L Q Q A G A R Y L V V W L L P D I G
721 GCGGCGGCCAATCGCCTGGGCGATGGCGTCCGCGCCCTGCAGCAGGCGCGGTGCGCGCTACCTGGTGGTCTGGCTGCTGCCGACATCGG
216 R T P A L A G S P Q Q A A S S A L S Q V Y N Q A L V A R L A
811 CCGGACGCGCGCGCTGGCGGCTCGCCGCAACAGGCGGCCAGCTCGGCCCTGAGCCAGGTCTACAACAGGCGTTGGTGGCCCCGCTGGC
246 G I D A E I I G L N V P Q L L A E V V A D P A R Y G L A T G
901 CGGCATCGATGCCGAGATCATCGGCCTGAACGTGCCGCACTGTGCTCGCCGAGGTGGTGGCCGATCCGGCGCGTACGGCCTGGCGACCGG
276 Q D L T G T C F S G D N C T R N T T Y G L G A A A A D P S Q
991 GCAGGTCTCACCGGGACCTGCTTCAGCGGTGACAACTGCACCCGCAACACCACTACGGCCTGGGCGCGCGGACGGGACCCGAGCCA
306 L L F N D R V H P T I S G R L I A D Y A Y S L L A A P W E
1081 GCTGCTGTTCAACGACCGCTCCACCCGACCATCAGCGGCCAGCGCTGATCGCCGACTACGCCTATTGCTGCTGGCCGCGCCCTGGGA
336 V T L L P E M A Q S T L R A H Q D G L H Q R W L G D W Y A W
1171 GGTACCCCTGTGCGGAGATGGCGCAGTCCACCCTGAGGGCGCACCCAGGATGGCCTGCACCAGCGCTGGCTGGGCGACTGGTATGCCTG
366 Q P V G R W Q G Y V D T S G R R L D V D D Q R S T A G G D G
1261 GCAGCGGTGGGGCGCTGGCAGGGTACGTGGATACCAGCGGGCGCGGCTGGACGTCGACGACGCGCAGTACCGCCGGGGCGATGG
396 N G Y G L D L G G S Y R L S D T W R L G V A A G V S R Q H L
1351 CAACGGCTATGGCTGGACCTGGGCGGAGTACCGGCTCAGCGACACCTGGCGCTCGGCGTGGCCGCGGGGTTTCGCGCCAGCACCT
426 E V G A A D S D Y R L N S Y L L S A F A Q Y Q G E R L W G D
1441 GGAGGTGGGTGCGGCGATTCCGACTATCGCCTGAACAGCTACCTGCTCAGCGCCTTCGCCCATAACAGGGCGAGCGGTCTGGGGCGA
456 L T A T G G R L D Y D D L S R R F V L G P T T R S E R G D T
1531 TCTACCGCCACCGGTGGCCGCTGGACTACGACGACCTGAGCCGTCGTTTCGTCTGGGCCGACACCCGACGCGAGCGCGCGGATAC
486 D G D L R A L S A R L G Y D L A A A S S P W H L S P Y L S A
1621 CGACGGCGACCTGCGCGCCCTCTCGGCGCGCTGGGGTATGACCTGGCGGCCGCCAGCAGTCCCTGGCACCTGAGTCCCTACCTGAGCGC
516 D Y A R V Q V D G Y R E R S S D A A A L A F A D Q S S T S R
1711 CGACTATGCCCGCTCCAGGTGGACGGCTATCGCGAGCGCAGCAGGATGCCCGCGCTGGCCTTTGCTGACCAGTCGAGTACCTCGCG
546 R L G L G L Q G R W Q F T P A T A L F A D V G R E R E F A D
1801 GCGGTGGGCTGGGGCTGCAGGGCCGCTGGCAATTACCCCGGCCACGGCGCTGTTCCCGATGTGGGCGCGAGGCGGAATTCGCGCA
576 G C T H D L T M S L N S V P G L D Y T L D G Y Q P D S S D R T C
1891 CGGTACCCATGACCTGACCATGAGCCTGAACAGCGTGCCGGGGCTGGACTACACCTGGACGGCTACCAGCCGACAGCGACCGCACCCG
606 L S L G V V Q R L T P E L T L R G G Y S Y A G A G D S H Q Q
1981 CCTGAGTCTGGGCTGGTGCAGCGGCTGACCCCGGAGTTGACCTCGCGGGTGGCTACAGCTATGCCGCGCGGGCGATAGCCATCAGCA
636 G V G L G L S L D F *
2071 GGGGGTAGGCCTGGGCCTTTGCTGGACTTCTAGCcgtagccaggccgggttgctgcccgcggcgagcacccttagaccggctgggacac
2161 gcccaggcgctgccgtgccagtcgcgcagctcgtcgaggcgccgcgggcgagatcaggaagccctgcaacaggctgcgacggcgctg
2251 ttgaggcgagagctgcttgagctcgtgtccacccttcggcgatgacctggcgctcgcgcgatggcagaaggcgacgaggcgctccag
2341 gctggcttggttgcgggttggtgtaggcgctcgacgatgagtcgatcgaacttgatggtgtcggggggtagtcggtgacctgctggat
2431 ggagggttagccacgccaagtcgtcgaatggccacgcgaagcccgcgcgcgagttcatgcagcagcgccagggtccgctcgttgag
2521 ttggcgggcgaaggtctcgggaattccagttccaggcgctccggatcgacgcccgtggtgacccaccgctcgacaggtagggcacgat
2611 atcgccctggtccagttgctcgcgagcagaggttgatggacagcaccaggtcgggcgcgagcactcctgcaacgccggatactccggca
2701 ggctggtccaccaccagcgatcgatccagcggaatgccctggcgctcgcgcatgggaatgaattccgcccgggacacggcgcccag

```

```

2791 gcggggggaattccagcgcaccagggtctcggcggagcggacctggccgtcgctgtcgacggccggcatgtagacgaggtggaattcctc
2881 gtccgggttcagctggcgcgcaaatgctcgtccagctgggtcatgcgccgctggcgcgcgccagctccggggtgaaatgcaccaggcgatt
2971 gccaccttcgcccttggcccgccagcgcaggtcggcctggcgagcaggtcggccaggtccagcggggcctcgggagcggccaggcc
3061 gaggtctgagggtcacggcatagcgcgttcgccgaccagggtgcccgttgctcgtagaggcgtgcaactggtcaccaggggccaggccatc
3151 ggcagccgcctggggcccgcgagcaggacggcgaattcatcaccggccaggcgggtgaagaccacctcgttgcttcgtagtagctgtc
3241 gatcaccccgctcgtctggcggccacggcctgcagcaactcgtcgccacctcatggcgaacctgtcgttcaccgacttgaagtgatc
3331 gaggcgaagtacagcagcgcgacggtccgacctccgacgctcgtcgagccagaggtcccccataactggaagtggcggcgattgc

```

B.1.8 N20 DNA Insert

```

1 aacggccgccagtgctggaattatatatatgggtacgcgcgttaagacaatatggctgttggttgccgcgacaactatgccgacctggag
91 cacaaggaggtgtggcagagtttcttggtgctggctatcgtactgttggtgttcgtcatctatcgctttaccaacgaaggtacgacttac
181 atctacgccatgctgtaattagcgcatactgatctgctacctcctgtggtgggtagagacgttaagcgcacctcagttatttctcaacag
271 cagagtcgtctcggcgagagcttgtcaccacggaggtagtggtatgcggaagattttgaacttccccagaccattcacgacaagatcggg
361 ccattgctgcagcagcattgcatagacacagctctatctgcagcagcatcttaccatcattcaacttgccaagctattggcactaac
451 cgctattatctcagtcagttatttctcccgtcagggtatcacatataatacctatgtcaacagtttgcgcatcaaccatttcaccaatctc
541 tatcgcaagtctgtcgccagtcagcgttcttcaccgccaggcaactggcttttgaaagtggctacagaaactatacgaccttcagtgcc
631 gtattcaaacagctgaagggcgagaccgttacagcgtggatgcataagacgccaataatccaagggtcgagtttcagaatcagcaaaa
721 tatcgcttcaaaaactcgcaaaaactgatttttctggtacttggcgcaatatattttggagattttgtttacctttgcaacataaccaca
1 M K T T T D Y S Q K A N W C K F P
811 tatttactagattatttatcaataaataaaataaaaacgATGAAAACAACACTACCGATTACTCCGAGAAGGCTAACTGGTGTAATTCCTCC
18 Q I T K D V D T F Y I F A T E Y I M G S F E E G A P D Y G T
901 CCAAAATTACCAAGGACGTAGATACGTTCTACATATTGCTACTGAGTATATCATGGGAAGCTTTGAAGAAGGTGCCCCAGATTATGGTAC
48 L D N Q E M I E G M N V E Y L V H A T T Y A D S T N V F A P
991 ACTCGACAATCAAGAGATGATAGAGGTATGAATGTTGAGTATTTAGTACATGCAACAACATATGCCGACTCAACCAATGTGTTGCTGCC
78 Y Y R Q T G L R F A G D I Y K R D G N F E A A L I G T P I D
1081 ATACTACCGCCAGACTGGTCTAAGGTTTGCTGGTGACATATATAAGAGAGATGGCAATTTGAGGCTGCGCTTATCGGTACGCCCATTTGA
108 D I M A A L D Y Y F E H Y N E G R P F I I A G H S Q G S A L
1171 TGACATCATGGCAGCCCTCGACTACTATTTGAGCACTACAACGAGGGCCGTCGCTTCATCATCGCAGGTACAGCCAAGGCTCCGCATT
138 V K V V L M K Y F K E H P D Y Y R R M V A A Y V I G Y S V T
1261 GGTCAAGGTGGTATTGATGAAATACTTCAAGGAACACCCAGACTATTACGGGCGCATGGTGGCTGCCTATGTTATAGTTACTCGGTAC
168 K E D L E N Y P Y L K F A T G E S D A G V I V S W N T E G P
1351 AAAAGAAGACCTTGAGAACTATCCGTATCTGAAATTCGCTACTGGCGAGAGTGATGCAGGGGTCAATTGTCAGTTGGAACACCGAAGTCC
198 K N V E E N T K T V V L L P N A I S I N P L N W K R D E T Y
1441 AAAGAATGTTGAGGAGAATACTAAGACCGTTGTGCTGCTACCAACGCTATCAGCATCAACCCGCTGAAGTGAAGCGCGACGAGACTTA
228 A P A S E N L G S L V V N E E T G E P E I G D L G A D A Q V
1531 TGCGCCCGCCAGCGAGAATCTGGGTCGCTTGTGGTGAACGAAGAACCCGAGAGCCTGAGATTGGTGATCTTGGTGGGACGCACAGGT
258 N L A R G T I V T H A K V V P M P E D A A K V A A E F F G P
1621 GAATCTTGCCCGAGGCACAATCGTGACGCACGCTAAAGTAGTCCCGATGCCGAGGATGCCGCAAGGTAGCCGAGAGTTCTTCGGACC
288 D G R H G E D Y T Y F Y N N I K D N V A K R I A T F K T N M
1711 CGATGGCCGTCATGGTGAAGACTATACATACTTCTACAACAATATCAAGGACAACGTTGCCAAGCGCATCGCCACCTTCAAGACGAATAT
318 R Q D
1801 GCGACAAGATatcgtcgggaccggtgcggaacctcgtggtctatggtaagattttcacttctgacaacgataaggtggtagaagcctttgc
1891 cgtgaaagacggtaagtatgtgtatgtggtgccaaggcaggagccgaagcctttattgaagcaggtgaagactgaagtggtggactacac
1981 cggaaagggacttgtagtgcaggttgccggtaatggtcacgcccactattcgataggtgttgccctgccaatcgtcggaacggttagccat
2071 tgggctgactctcgaccagtttatggcagaagctgtccccgctgcggtcaagaaggctcgcgagacaggagcgacgtccatcttcggtt
2161 cggttggaattacatcaccttcattggagaatatgccaccgctcagcaactggatgccatctgcagtgatattcctatatattttgcca
2251 cgacgaaggtcacaaggggctggcaaacaccctgatgctggtcaaggccggaattatgaaagccgacggcacggtgctcaagaaggataa
2341 ggatatccgtggcggtgagatcgtgatgggtgctgacggtacgccaaccgggttcctgaaagaacaggcaggcacttatacacgttcctt
2431 cctcgacaccgagaaactctatcccgttgacgtcgcaaaaattgtgtctcaaaggtccaggaacaacttctgtcagagggttatcacgat
2521 gtatatcgatggttggggcaactatttcttcaaccac

```

B.1.9 N26 DNA Insert

```

1 ttgacaatcgtgctgtaggctgcattgtgtgcgccgaggtgcgcaagccatgggttcaacattgtcaacactgcgctcaattgtgcatg

```

```

91 cgttgccgggttgcccacgccttatgctgcccgtgattaattccagcctgaagccttttgaggctgatggcagtatcaagaatcccgcag
1      L S S L P A A R A L P P A S R R P L I
181 tggccgctcagctgaatattatggccggtcaggTTGTCGAGTTTGCCCGCGCGCGCCTTGCCGCCAGCGTCGAGGCGTCCGTTAATT
20 K T G V S N M A F E E S F I T V D G C R T R I R R G G K G P
271 AAAACAGGAGTCAGCAATATGGCGTTTGAAGAGTCTTTTATTACCGTTGATGGTTGCCGAACCCGTATCCGGCGGGGTGGCAAAGGCCA
50 T V L Y L H G A N G A P M I Q P F M E V L A Q D Y D L I V P
361 ACGGTGTTGATTTGTCATGGTGCAAATGGTGCGCAATGATTGAGCCTTTCATGGAAGTGCTGGCGCAAGATTATGATTGATCGTGCCC
80 E H P G F G M S D E P E W L D S M Q D L A Y F Y L D L L D H
451 GAGCACCTGTTTTGGTATGTCGGATGAGCCGAATGGCTCGACAGCATGCAAGACCTGGCATATTTTATCTCGATTGCTCGACCAT
110 M K L D S V H V V G S S M G G W L G M E M G I R E P R R I K
541 ATGAAGTAGACAGCGTGCATGTGGTGGGTAGTTCGATGGGTGGCTGGCTCGGTATGGAATGGGTATTCGTGAGCCCCGTGCTATTAAA
140 S L T L V G T A G V R V P G I L P G D I F L W D A E T A A R
631 TCGCTTACTTTGGTGGGTACGGCCGGTGTGCGTGTCCGGGTATTTTGCCAGGGGATATTTTCTCTGGGACGCAGAAACGGCAGCCCGC
170 N T F F N Q D I A Q K V L S M A P K T E E A Q D I M L K N R
721 AATACGTTCTTCAACCAAGACATTGCGCAGAAAGTGCTGTCGATGGCACCCAAAACCGAGGAAGCGCAGGACATCATGCTCAAGAATCGG
200 E T V A R L A W Q P R L F D L N L P K W L H R I Q A P V K L
811 GAAACGGTTGCAAGGCTGGCTTGGCAACCGCGCTTGTTCGATCTTAACTTGCCCAAGTGCGTGCATTCAGGCGCCGGTCAAGCTG
230 I W G E Q D K I M P L A V G E A L Q P K L P N A T L Q V F K
901 ATATGGGGTGAGCAAGACAAGATTATGCCTTTAGCGGTGGCGAGGCGTTCGAGCCAAACTGCCGAATGCGAGCTGCGAGGTGTTCAAA
260 N C G H L P Q V E F P N E F S A S V K Q F I E G V K *
991 AACTGTGGTCAATTTACCGCAGGTTGAGTTTCCCAACGAGTTGAGCGCGTCCGTTAAGCAATTTATTGAAGGGGTTAAATAGccatgaatg
1081 tcacgcttttccatttgatgtcgtatgcccgcacttgactttgaggccacaaaggaatatgaaaccgtttgggatgaaattgccaacaagt
1171 tctatgaccccggtgaaaggccacaagctttataaccgttatctcgacgagctcgagtatgcagaaacgc

```

B.1.10 N33 DNA Insert

```

1 gatggagcagatgggtctgcgcaaggcgcatctgaccaacatgatccccgacggcaaggccgtatccgcctggaatacaccatcccggc
91 gcgtggcctgatcggcttccgcaacaacttctgaccctgacctcggcagcgcatcctgacctgaccttcagccactacggcccgat
181 caaggccggcgaagtacgacaaccgcagaacggcgctgctggtgtccatggccaccgggtaccgctgacctactcgtggaaacctgca
271 gagccgcggaagctgttccctggctccggcgacgagatctacgaaggccagctgtgcgccatcaacagccgtgacaacgacctggtgct
361 caaccccaaccaaggcgaagaagctggacaacatgcgcgcctcgggcaaggacgaagtcacgacctggtgcccgcgatcaagttcacct
451 cgagcaggcgctggaattcatcgccgacgacgagctggtggaagtacgccgaagtcacccgctgcgcaagaagtacctgaacgagaa
541 cgaccgcaagcgcttcgagcgttccaaggtctgatcgcttccggtaacgcaagaaagcgctccggcgctttttgtgtcccgcgtgc
631 cgccatcgacagcgctctatcgcgccatggcggtccgctgctggtcgctccgacgatcacagcgggcccgtatgagcggcccatggc
721 cgcgatagatcgctagccatgggatcacggaatatcctggcgcgccctccctcacccggccctctccagggggagagggggaaaaccg
811 ctcccgtagctaggttgggttgagcgagcgaagcccagccatttcaccgccaggccgctccgggagaggtggaaccgctcgcgtaggtt
901 gggttgagcgagcgaagcccaacagccccgccccctggcgctagatggcgcgaccttcaggctcgcgcgagcgacgctgccgtcccgt
991 cgccttgcaacccctctggcccaggcctctcagcgctgcaacgaacgtctcgtactttttgttcacaaccagatacaacgtcataaaaa
1      M T F S L R Q R L P
1081 ccgtctagcttcagacgacgccccaacgcgtgctccgataacaacaatcaagggtgcaggcATGACATTTTCTCTCAGGCAACCGCTGCC
11      L A I A L A T A L A A T V Q A A P N P Y S G F T V F G D S L
1171 GCTGGCCATCGCCCTGGCGAGCGGCTCGCGCAACCGTGACAGCGGCTCCCAATCCCTACAGTGGCTTACCGTGTTCGGCGACAGCCT
41      L D A G Q F P D T G V T G A S L R F T N R V G P G Y S A A G
1261 GCTTGATGCCGGGCAATTTCCCGATACCGGGGTGACCGGCGCCAGCCTCAGATTACCAACCGGGTCGGCCCCGGGTACAGCGCCGCGG
71      G A V T G P V S S I L L G Q Q L G F D A R T L D A S T S V I
1351 CGGCGCGGTGACCGGACCGGTGTCGTCGATCCTGCTCGGCCAGCAACTGGGTTTCGACGCCCGGACCCTGGACGCCTCGACCTCGGTGAT
101      N R L L G L A A G D N W A T G G Y T T A Q I R D S I T A A N
1441 CAATCGGTGCTGGGCCTCGCCCGCGGCGACAACCTGGGCGACCGGTGGCTACACCAGGCGCAGATTGCGGATTCCATCACCGCCGCCAA
131      G S V V A A N G L T L R S R D G Y L P G L A S Q G L R L D P
1531 CGGCTCGGTGGTGGCGGCAATGGCCTGACCCTGCGCAGCCGCGACGGCTACCTGCCCCGACTCGCCAGCCAGGGGCTGCGCCTGGATCC
161      N T L F Y I S G G G N D F L Q G L V T S P A S A A A A A N R
1621 CAACACCTGTTCTATATCAGCGGGGGCGGTAACGACTTCTCCAGGGCTGGTACCTCGCCGGCAGCGCCGCGGCGGCGGCAATCG
191      L G D G V A A L Q Q A G A R Y L V V W L L P D I G R T P A L
1711 CCTGGGCGATGGCTCGCCGCCCTGCAGCAGGCCGCTGCGCGCTACCTGGTGGTCTGGGTGCTGCCGACATCGGCGGACGCCGGCGCT
221      A G S P Q Q A A S S A L S Q V Y N Q A L V A R L A G I D A E

```

```

1801 GGCCGGCTCGCCGCAACAGGCGGCCAGCTCGGCCCTGAGCCAGGTCTACAACCAGGCGTTGGTGGCCCGCTGGCCGGCATCGATGCCGA
251 I I G L N V P Q L L A E V V A D P A R Y G L A T G Q D L T G
1891 GATCATCGGCCTGAACGTGCCGAGTTGCTCGCCGAGGTGGTGGCCGATCCGGCGCGCTACGGCCTGGCGACCGGGCAGGATCTCACCGG
281 T C F S G D N C T R N T T Y G L G A A A A D P S Q L L F N D
1981 GACCTGCTTCAGCGGTGACAACTGCACCCGCAACACCACCTACGGCCTGGGCGCCGCGGACGGACCCGAGCCAGCTGCTGTTCAACGA
311 R V H P T I S G Q R L I A D Y A Y S L L A A P W E V T L L P
2071 CCGCGTCCACCCGACCATCAGCGGCCAGCGGCTGATCGCCGACTACGCCTATTGCTGCTGGCCGCGCCCTGGGAGGTACCCCTGCTGCC
341 E M A Q S T L R A H Q D G L H Q R W L G D W Y A W Q P V G R
2161 GGAGATGGCGCAGTCCACCCTGAGGGCGCACCCAGGATGGCTGCACAGCGCTGGCTGGGCGACTGGTATGCCTGGCAGCCGGTGGGGCG
371 W Q G Y V D T S G R R L D V D D Q R S T A G G D G N G Y G L
2251 CTGGCAGGGCTACGTGGATACAGCGGGCGCCGGCTGGACGTGCAGCAGCAGTACCGCCGGGGCGATGGCAACGGTATGGCCCT
401 D L G G S Y R L S D T W R L G V A A G V S R Q H L E V G A A
2341 GGACCTGGGCGGACGTACCGGCTCAGCGACACCTGGCGCTCGGCGTGGCCGCGGGGTTTCGCGCCAGCACCTGGAGGTGGGTGCGGC
431 D S D Y R L N S Y L L S A F A Q Y Q G E R L W G D L T A T G
2431 CGATTCCGACTATCGCCTGAACAGTACCTGCTCAGCGCCTTCGCCAATACCAGGCGGAGCGGCTCTGGGGCGATCTACCGCCACCGG
461 G R L D Y D D L S R R F V L G P T T R S E R G D T D G D L R
2521 TGGCCGGCTGGACTACGACGACCTGAGCCGTCGTTTCGCTCGTGGCCCGACCAACCGCAGCGAGCGCGGATACCGACGGCGACCTGCG
491 A L S A R L G Y D L A A A S S P W H L S P Y L S A D Y A R V
2611 CGCCCTCTCGGCGCGCTGGGTATGACCTGGCGGGCGCCAGCAGTCCCTGGCACCTGAGTCCCTACCTGAGCGCGGACTATGCCCGCGT
521 Q V D G Y R E R S S D A A A L A F A D Q S S T S R R L G L G
2701 CCAGGTGGACGGCTATCGCGAGCGCAGCAGCGATGCCGCGCGCTGGCCTTTGCTGACCACTGAGTACCTCGCGCGGGCTGGGCGTGGG
551 L Q G R W Q F T P A T A L F A D V G R E R E F A D G T H D L
2791 GCTGCAGGGCGCTGGCAATTACCCCGGCCACGGCGCTGTTTCGCCGATGTGGCCCGGAGCGGGAATTCGCCGACGGTACCCATGACCT
581 T M S L N S V P G L D Y T L D G Y Q P D S D R T R L S L G V
2881 GACCATGAGCCTGAACAGCGTGC CGGGGCTGGACTACACCTGGACGGCTACCAGCCGACAGCGACCGCACCCGCTGAGTCTGGGCGT
611 V Q R L T P E L T L R G G Y S Y A G A G D S H Q Q G V G L G
2971 GGTGCAGCGGTGACCCCGAGTTGACCTGCGCGGTGGCTACAGCTATGCCGCGCGGGCGATAGCCATCAGCAGGGGGTAGGCTGGG
641 L S L D F *
3061 CCTTTCGCTGGACTTCTAGccgtagccagccgggttgctgccccgcggcagcaccacctagaccggctgggatcgccagcgctgccc
3151 ctgccagtcgcgagctcgtcgaggcgccgcgggcgagcatcaggaagccctgcaacaggtcgcgagccggcctgttcgagcgagagctg
3241 cttgagctcgctgtccacccttcgcgcatgacctggcgctcgcgcatggcagaaggcgaggggctccca

```

B.1.11 RR11 DNA Insert

```

1 M D E A K I Q
1 acaaagagggtgaaaaaacgccgttctactttatttcacggcggtggttatcattttggcaatgtgagtATGGATGAAGCGAAGATAC
8 G V A D G C G T T V I A P D Y T L S L D P S Y K Y P M E L E
91 AGGGTGTGGCAGATGGGTGCGGCACCAAGTGTAGCTCCTGACTATACGCTTTCGCTGGATCCTTCATACAAATATCCCATTGGAGCTTG
38 Q V Y A G L L Y A Y E H A D E L N I D A D N I V I E G E S A
181 AACAGGTATATGCCGACTCCTTTATGCTTATGAGCATGCTGACGAATTAATATCGATGCGGATAACATTGTTCATTGAGGGAGAAAGCG
68 G G G L T A R L A L Y N K D K G K V P L K G Q V L I Y P M L
271 CCGGTGGTGGACTGACTGCACGCCTGGCTCTTTACAACAAGGACAAGGAAAGGTTCCCTGAAGGGTGCAAGTGTGATATACCCATGCG
98 D Y R T G G E K D I Y K N E Y A G D C I W T K E N N I F G W
361 TTGACTACCGTACCGGCGCGAAAAAGACATTTACAAGAATGAATATGCCGGGGATTGTATCTGGACCAAGGAGAATAACATCTTCGGAT
128 G K L V E G Q E K K L T D E E M I Y F S P A V A T A E Q L K
451 GGGGAAACTGGTTGAAGGTCAGGAAAAGAACTTACCGACGAGGAAATGATCTATTTTACCCGCACTGGCAACTGCTGAGCAACTGA
158 G L P E T F V I V G S L D L F C D E D I S Y A Q K L M E A G
541 AGGGATTGCCAGAGACTTTTCGTGATTGTGCGGACGCTTGATCTTTTCTGCGATGAAGACATTTCTATGCCAGAGCTTATGGAGGCTG
188 V F T E L H V E P G V P H A Y E Y L E W T P Q A H R F I E M
631 GTGTTTTACAGAACTCCATGTAGAACCCGAGTTCTCAGCCTTATGAGTACTTAGAGTGGACACCTCAGGCGCATAGGTTTATTGAAA
218 R N H A T A R M L G A E K D V K E S A E A K A F R E L L S K
721 TGAGAAACCATGCGACAGCAGGATGCTTGGAGCTGAAAAAGAGCTAAAAGAGTCTGCTGAAGCAAAGGCTTTCAGAGAACTTTGTCAA
248 Y N I E Q * M S E N M N R P L D Y S Q K
811 AGTATAACATTGAGCAGTAAtacagcctaattaaaaccttataatacaATGAGTGAAAATATGAATCGGCCATTAGATTATCTCAGAAG
268 A N W Y Q I P E V T K E F D T F Y V Y A T E Y I L S S M E E

```

901 GCCAATTGGTATCAGATTCCAGAAGTTACAAAGGAATTCGACACATTTTATGTATATGCAACCGAGTACATCCTGAGTAGCATGGAGGAG
298 G A P D Y A D M E N A E M L E G A A A E Y M L H A T A Y A D
991 GGTGCTCCGGATTATGCTGATATGGAAAATGCCGAAATGCTGGAGGGTCTGCTGCAGAATATATGTTGCACGCCACCGCATACGCAGAT
328 S T N V F M P Y Y R Q V G L R Y A G V V W K R D G I F D A S
1081 TCCACCAACGTGTTTCATGCCTTACTACCGTCAAGTAGGCTTGCGTTATGCCGGGGTCTGCTGGAAGAGGGACGGAATCTTCGATGCCTCC
358 V A G M P Y G D I V A A L D Y Y F E H Y N N G R P F I I A G
1171 GTTGCTGGTATGCCTTATGGCGACATTGTTGCCGCATTAGACTATTACTTCGAGCACTACAACAACGGCGGCCATTTCATCATCGCTGGC
388 H S Q G S A I I K M V L K K Y F V E H P D Y Y K R M I A A Y
1261 CATAGTCAGGGATCGGCTATTATTAAGATGGTGCTAAAAAATATTTTGTGGAGCACCCCGATTACTACAAGCGTATGATTGCCGCTTAC
418 P I G Y A F T K D E F K T Y P H M K F A T G E C D T G V I I
1351 CCGATTGGCTACGCTTTACGAAGGACGAGTTCAAGACATACCCACATATGAAGTTTGCCACTGGTGAGTGCGATACGGGTGTTATCATC
448 T W N T E G P K N R E V N A D T C V L Q P G S M S I N P L N
1441 ACCTGGAATACCGAAGGACCTAAGAACAGGGAAGTAAACGCTGATACGTGTGTGCTGCAGCCGGGTTCTATGAGTATAAACCCGCTTAAC
478 W K L D D T Y A P A S M N L G S L F P N K D T G K L E I Q D
1531 TGGAAACTGGACGACACTTATGCTCCGGCCAGCATGAATCTGGGGTCTCTTTTCCGAACAAGATACCGGCAAACTGGAGATTCAAGAC
508 L G A D A Q V F P D R G V V V T H A M G E E M T E E T A K V
1621 CTGGGTGCGGACGACAGGTGTTCCCTGACCGCGGAGTGGTAGTTACCCATGCCATGGGTGAGGAAATGACCGAGGAGACGCCAAAGTG
538 A A E F F G P D G R H G E D Y V L F Y C N I K D N V A K R I
1711 GCAGCCGAATTCTTCGGCCCTGATGGCCGTCACGGCGAAGATTACGTACTTTTCTATTGCAACATTAAGGATAATGTGGCTAAACGTATC
568 S A Y M A N R K *
1801 TCGGCCTATATGGCAAACCGCAAGTAGcagaacaattcattatattaaaaagattataatgaacaaaaatgtatttataaatataataa
1891 atattttatggtacaagattttatgttttgatgattacagcagctgtagttagtattctttttaagttgttaaagcagccagtggtgctg
1981 ggttatattgttcaggtgtcttggttggtccatatctgtttggcaaaagccttgatccggacaacgtgaagcaatggggcgacatc
2071 ggtgtgctgtttgtattgttctgcatcggtctggagttccgcctgaagaacctcatcagcagcggaaggtggcagcaataggtgcgttg
2161 acaatcatcttggttatgatgctcttaggttacgtcgtgggagtggtgatgcta

Table B.2 Blast Top Hits.

Name	Top Hit [†]	Query Cover	Identity	GenBank ID	Submitted
N1ORF4	hypothetical protein	99%	88%	WP_093730338.1	07/2017
N1ORF5	hypothetical protein	99%	90%	WP_090815217.1	07/2017
N2	autotransporter domain-containing protein	99%	91%	WP_081564916.1	12/2017
N4	hypothetical protein	99%	99%	EKE23468.1	09/2012
N7	alpha/beta hydrolase	96%	52%	WP_091149153.1	08/2017
N11	alpha/beta fold hydrolaseb	99%	98%	WP_089991751.1	06/2018
N13	lipase	99%	97%	WP_025328824.1	06/2014
N16	carboxylesterase/lipase family protein	99%	62%	WP_051543757.1	05/2017
N18	autotransporter domain-containing protein	99%	93%	WP_059316517.1	20/2017
N20	DUF3089 domain-containing protein	96%	71%	WP_044931439.1	07/2017
N26	alpha/beta fold hydrolase	90%	96%	WP_114420284.1	07/2018
RR11ORF1	steryl acetyl hydrolase	90%	81%	WP_029200816.1	11/2017
RR11ORF2	EstGK1	98%	85%	ADE28719.1	04/2010

[†] as of 09/2018

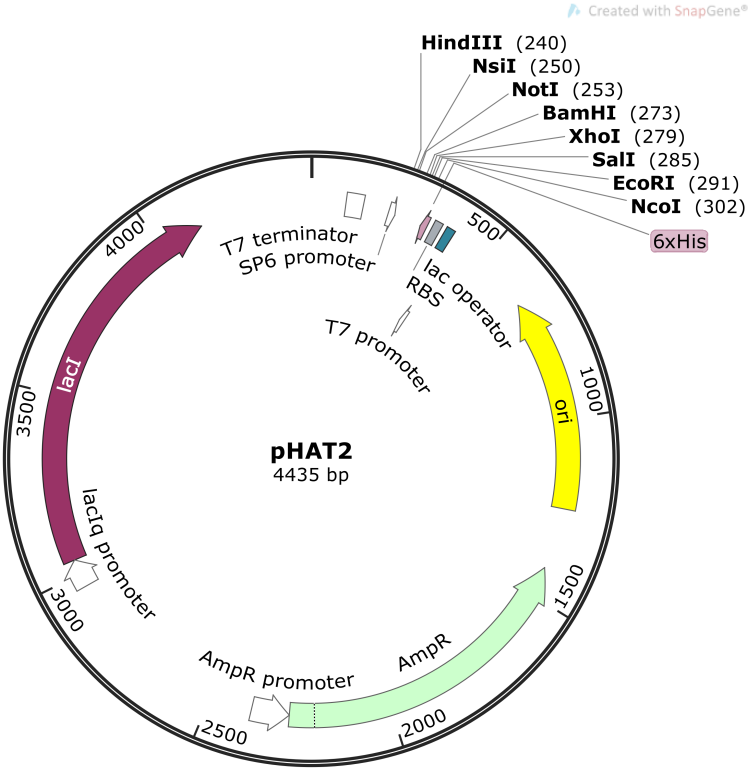


Fig. B.2 The Vector Map of pHAT2.

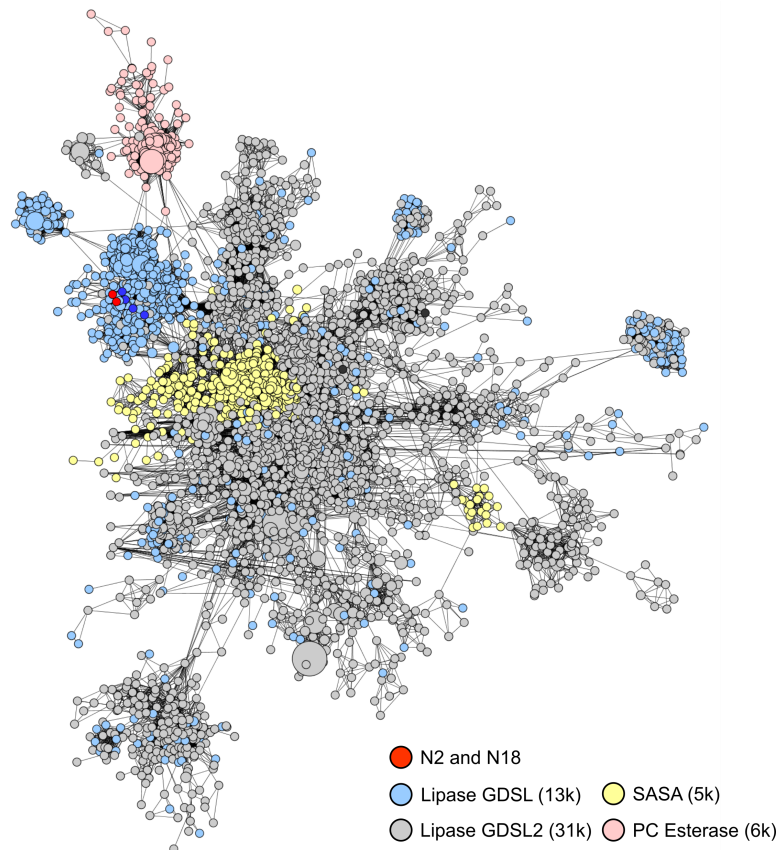


Fig. B.3 Sequence similarity network of the SGNH superfamily (Pfam CL0264) containing hits N2 and N18. They are members of the Lipase GDSL 2 family, where they are located centrally near other functionally characterised GDSL Lipases (darker blue). The network built using the four largest families in CL0264 (62,000 sequences clustered into 4,000 nodes representing 90% of the clan's sequences), connecting lines represent e-values $<10^{-9}$, node size is proportional to the number of sequences represented (range 1 to 1165).

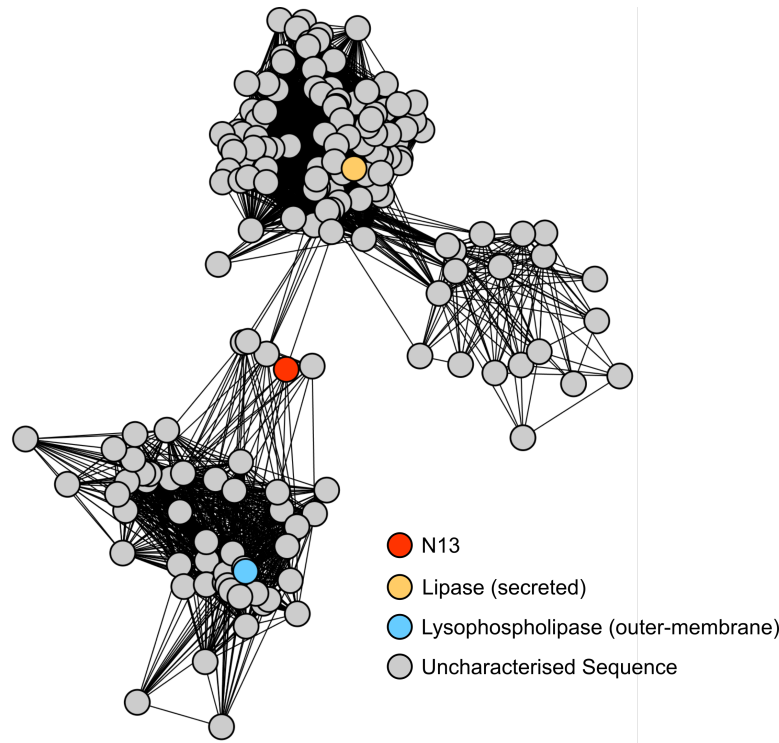


Fig. B.4 Sequence similarity network for N13, which is a member of the Lipase_bact_N family. N13 (red) sits between to clusters which each contain one characterised enzyme, a lipase from *A. hydrophila* (LIPE_AERHY, yellow) and a lysophospholipase of *V. cholerae* (VOLA_VIBCH, blue). Built using Pfam family PF12262 (230 sequences), connecting lines correspond to e-values $<10^{-100}$, nodes represent one sequence each.

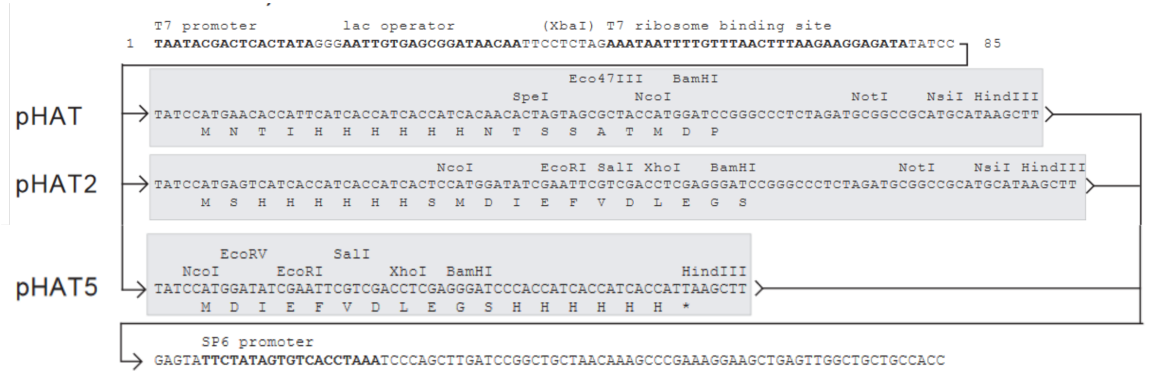


Fig. B.5 The vectors pHAT, pHAT2, and pHAT5 differ only in their multiple cloning site. pHAT and pHAT2 can be used to add an N-terminal His-6-tag, pHAT5 can be used to add a C-terminal His-tag.

Table B.3 Cloning sites for each hit in vector pHAT as well as expression strain and temperature.

Name	N-ter	C-ter	Gene (kbp)	MW (kDa)	Strain	Temperature °C
N1ORF4	SpeI	HindIII	1.6	61.4	BL21(DE3)	37
N1ORF5	SpeI	NotI	1.5	58.3	BL21(DE3)	20
N2Lip	SpeI	HindIII	1.0	36.2	BL21(DE3)	15
N4ΔSP	SpeI	NotI	1.2	44.0	SHuffle T7 Express	15
N7	SpeI	NotI	1.0	37.1	BL21(DE3)	20
N11	SpeI	HindIII	0.8	31.5	SHuffle T7 Express	20
N13	SpeI	HindIII	2.4	84.2	BL21(DE3)	20
N16	SpeI	HindIII	1.5	58.4	BL21(DE3)	20
N18	SpeI	HindIII	1.9	68.1		
N20	NcoI	NotI	0.9	37.8	BL21(DE3)	20
N26	SpeI	HindIII	0.9	33.4	BL21(DE3)	20
RR11ORF1	SpeI	Eco47III	0.9	32.9	BL21(DE3)	20
RR11ORF2	SpeI	HindIII	1.0	36.4	BL21(DE3)	37

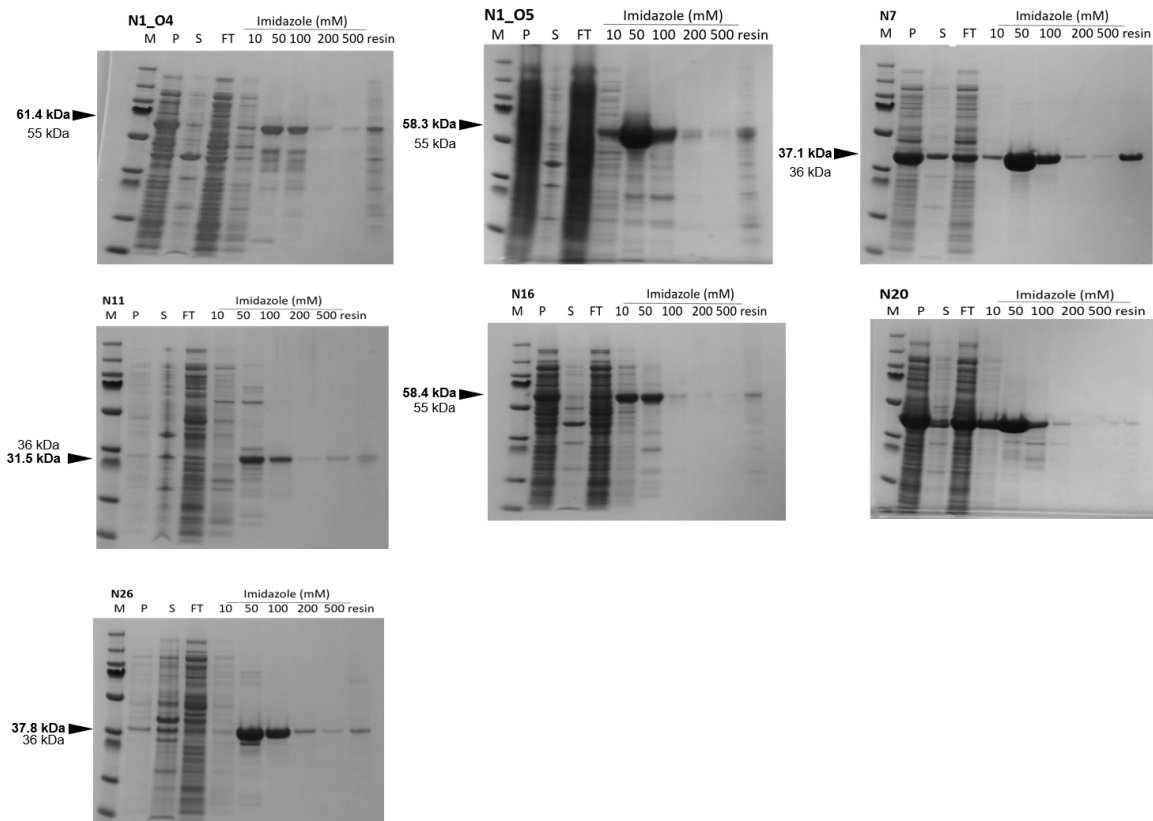


Fig. B.6 SDS-PAGE of protein purification.

Table B.4 Michaelis-Menten parameters of the metagenomic hits with pNP – O – C(=O) – (CH₂)_n – CH₃..

Enzyme	n	K _m / M	v _{max} /(mU/min)	k _{cat} / s ⁻¹	k _{cat} /K _m /(Ms) ⁻¹
N1ORF4	0	7.45E-04 ± 8.27E-05	3.38E+02 ± 1.34E+01	1.10E+01	1.47E+04
	1	1.29E-03 ± 8.64E-04	9.58E+02 ± 4.34E+02	3.10E+01	2.41E+04
	2	9.65E-04 ± 3.96E-04	4.77E+02 ± 1.26E+02	1.55E+01	1.60E+04
	3	9.71E-04 ± 1.00E-04	4.50E+02 ± 1.78E+01	1.46E+01	1.50E+04
	4	9.48E-04 ± 1.96E-04	8.00E+02 ± 6.33E+01	1.64E+00	1.73E+03
	5	1.41E-03 ± 8.52E-04	2.43E+02 ± 9.97E+01	2.49E-01	1.77E+02
	6	6.62E-04 ± 4.02E-05	1.31E+02 ± 4.53E+00	3.32E-02	5.01E+01
	7	3.86E-03 ± 1.04E+03	8.69E+01 ± 1.68E+01	6.32E-03	1.64E+00
N1ORF5	0	7.74E-04 ± 3.16E-05	1.54E+02 ± 2.27E+00	6.06E-01	7.83E+02
	1	1.20E-03 ± 4.29E-04	7.09E+02 ± 1.65E+02	2.79E+00	2.33E+03
	2	8.06E-04 ± 2.48E-04	4.56E+02 ± 8.44E+01	1.80E+00	2.23E+02
	3	1.69E-03 ± 5.89E-04	1.03E+03 ± 2.73E+02	1.63E+01	9.65E+03
	4	1.78E-03 ± 1.12E-03	6.13E+02 ± 2.96E+02	1.03E-01	5.81E+01
	5	3.46E-04 ± 1.16E-04	3.77E+01 ± 3.71E+00	7.00E-04	2.17E+00
N2	0	8.45E-03 ± 4.06E-01	7.38E+02 ± 2.61E+01	4.54E+00	5.37E+02
	1	1.46E-03 ± 2.52E-01	3.28E+02 ± 2.46E+01	1.01E+00	6.89E+02
	2	2.44E-04 ± 3.15E-02	4.72E+01 ± 1.75E+00	1.45E-01	5.97E+02
N4	0	6.23E-03 ± 1.22E+00	2.03E+03 ± 2.72E+02	6.08E+00	9.77E+02
	1	1.54E-03 ± 3.22E-01	1.36E+03 ± 1.26E+02	4.08E+00	2.65E+03
	2	4.28E-04 ± 6.08E-02	9.54E+02 ± 4.18E+01	2.87E+00	6.70E+03
	3	3.53E-04 ± 3.95E-02	6.74E+02 ± 2.22E+01	2.86E+00	8.11E+03
	4	2.11E-03 ± 8.21E-02	1.28E+03 ± 2.45E+01	5.43E+00	2.58E+03
	5	1.01E-03 ± 1.59E-01	2.88E+02 ± 1.77E+01	1.36E+00	1.35E+03
	6	1.32E-03 ± 4.62E-01	6.64E+01 ± 9.81E+00	3.95E-01	2.99E+02
N7	0	1.37E-03 ± 1.63E-04	3.38E+02 ± 1.72E+01	3.33E+00	2.43E+03
	1	3.14E-03 ± 3.27E-04	6.16E+02 ± 3.58E+01	6.06E+00	1.93E+03
	2	2.81E-03 ± 2.47E-03	1.05E+03 ± 7.41E+02	1.04E+01	3.69E+03
	3	5.11E-04 ± 3.23E-05	2.69E+02 ± 5.48E+00	6.62E-01	1.29E+03
	4	4.98E-04 ± 5.38E-05	6.48E+02 ± 2.24E+01	1.60E-01	3.20E+02
	5	8.79E-04 ± 1.88E-04	1.07E+02 ± 8.58E+00	2.63E-02	3.00E+01
	6	7.74E-04 ± 5.70E-05	1.08E+02 ± 2.87E+00	9.30E-03	1.20E+01
	7	3.64E-04 ± 2.97E-05	3.51E+01 ± 8.50E-01	3.00E-03	8.33E+00
N11	0	5.29E-03 ± 5.31E-03	2.02E+03 ± 1.57E+03	9.96E-01	1.88E+02
	1	2.67E-03 ± 5.08E-05	7.11E+02 ± 7.16E+00	3.50E-01	1.31E+02
	2	1.26E-03 ± 1.24E-04	5.08E+02 ± 2.08E+01	2.50E-01	1.98E+02
	3	1.26E-03 ± 5.96E-05	2.92E+02 ± 5.75E+00	1.44E-01	1.14E+02
	4	7.44E-04 ± 5.18E-05	9.45E+01 ± 2.34E+00	3.10E-02	4.17E+01

Table B.5 Michaelis-Menten parameters of the metagenomic hits with pNP – O – C(=O) – (CH₂)_n – CH₃.

Enzyme	n	K _m / M	v _{max} /(mU/min)	k _{cat} / s ⁻¹	k _{cat} /K _m /(Ms) ⁻¹
N13	0	3.65E-03 ± 2.92E-01	2.76E+02 ± 1.29E+01	7.31E+00	2.00E+03
	1	1.81E-03 ± 2.72E-01	6.18E+02 ± 4.33E+01	6.18E+00	3.41E+03
	2	7.51E-04 ± 1.32E-01	6.23E+02 ± 3.90E+01	6.24E+00	8.30E+03
	3	7.15E-04 ± 8.85E-02	4.17E+02 ± 1.82E+01	4.17E+00	5.84E+03
	4	1.27E-03 ± 8.77E-02	5.56E+02 ± 1.61E+01	5.56E+00	4.39E+03
N16	0	3.56E-02 ± 1.85E-02	1.18E+04 ± 5.63E+03	3.64E+00	1.02E+02
	1	1.12E-02 ± 2.92E-03	3.49E+03 ± 7.11E+02	1.72E+01	1.53E+03
	2	4.61E-03 ± 1.09E-03	6.99E+02 ± 5.48E+02	3.44E+00	7.47E+02
	3	1.71E-03 ± 6.20E-04	3.67E+02 ± 9.72E+01	4.52E-01	2.64E+02
	4	7.67E-04 ± 1.27E-04	1.52E+02 ± 1.51E+01	4.69E-02	6.11E+01
N20	0	1.33E-03 ± 8.08E-05	1.57E+02 ± 3.40E+00	6.19E+01	1.63E+04
	1	2.26E-03 ± 2.51E-04	2.60E+02 ± 2.12E+01	6.39E+02	2.83E+05
	2	1.43E-03 ± 6.54E-04	6.57E+02 ± 2.10E+02	2.59E+02	1.81E+05
	3	1.17E-03 ± 2.52E-04	1.89E+03 ± 2.60E+02	3.71E+02	3.17E+05
	4	1.55E-03 ± 2.01E-04	3.21E+02 ± 2.74E+01	3.16E+01	2.03E+04
	5	2.07E-03 ± 1.10E-03	1.03E+03 ± 3.52E+02	5.01E-01	2.45E+02
	6	4.43E-03 ± 5.06E-04	4.36E+02 ± 2.49E+01	4.29E-02	9.70E+00
	7	4.85E-03 ± 1.18E-03	1.63E+02 ± 2.04E+01	1.60E-02	3.31E+00
N26	0	6.90E-04 ± 1.57E-04	5.11E+02 ± 5.87E+01	5.03E+00	7.29E+03
	1	7.50E-04 ± 4.68E-04	3.44E+02 ± 1.55E+02	3.39E+00	4.52E+03
	2	8.70E-04 ± 1.13E-04	5.07E+02 ± 2.46E+01	1.25E+00	1.43E+03
	3	7.46E-04 ± 1.41E-04	4.13E+02 ± 2.79E+01	1.02E+00	1.36E+03
	4	8.27E-04 ± 1.28E-04	2.84E+02 ± 1.62E+01	7.00E-01	8.26E+02
	5	1.87E-03 ± 1.26E-03	2.80E+02 ± 1.48E+02	6.88E-01	3.68E+02
	6	7.97E-04 ± 5.95E-05	1.80E+02 ± 4.89E+00	1.11E-01	1.39E+02
	7	1.26E-03 ± 3.81E-04	8.35E+01 ± 1.05E+01	5.14E-02	4.09E+01
RR11ORF1	0	2.18E-03 ± 1.55E-01	1.11E+03 ± 3.92E+01	1.09E+00	5.01E+02
	1	2.86E-03 ± 3.31E-01	1.36E+03 ± 8.50E+01	6.67E-01	2.33E+02
	2	4.62E-04 ± 1.71E-01	4.39E+02 ± 5.09E+01	2.16E-01	4.68E+02
	3	4.84E-04 ± 1.20E-01	3.50E+02 ± 2.75E+01	1.72E-01	3.56E+02
	4	7.14E-04 ± 5.93E-02	1.58E+02 ± 4.61E+00	4.13E-02	5.78E+01
RR11ORF2	0	3.02E-02 ± 4.41E+00	5.02E+03 ± 6.60E+02	3.96E+02	1.31E+04
	1	2.23E-03 ± 5.99E-01	5.83E+02 ± 7.83E+01	4.63E+01	2.08E+04
	2	7.29E-04 ± 2.39E-01	4.54E+02 ± 5.28E+01	3.58E+01	4.91E+04
	3	3.79E-04 ± 4.16E-02	1.48E+02 ± 4.86E+00	1.17E+01	3.08E+04
	4	2.00E-03 ± 1.99E-01	2.79E+02 ± 1.34E+01	5.49E+00	2.74E+03
	5	2.84E-03 ± 4.16E-01	1.26E+02 ± 9.98E+00	7.83E-01	2.76E+02
	6	1.54E-03 ± 2.04E-01	3.73E+02 ± 2.20E+01	3.96E-02	2.57E+01
	7	7.42E-03 ± 2.17E+00	3.21E+02 ± 6.70E+01	1.70E-02	2.30E+00

Appendix C

Supplementary Data for Chapter 4

C.1 Amino-acid sequence of HG3.17

MAEAAQSIDQLIKARGKVYFGVATDQNRLLTGKNAAIKADFGMVWPEES
MQWDATEPSQGNFNFAGADYLVNWAQQNGKLIGAGCLVWHNFLPSWVSSI
TDKNTLINVMKNHITTLTRYKKGKIRTWDVVGEAFNEDGSLRQNVFLNVI
GEDYIPIAFQTARAADPNKLYIMDYNLDSASYPKTQAIVNRVKQWRAAG
VPIDGIGSQMHLASAGGAGVLQALPLLASAGTPEVSILMLDVAGASPTDY
VNVVNACLVNQSCVGITVMGVADPDFAFASSTPLLFDGNFNPKPAYNAIV
QNLQQGSLEGSHHHHHH

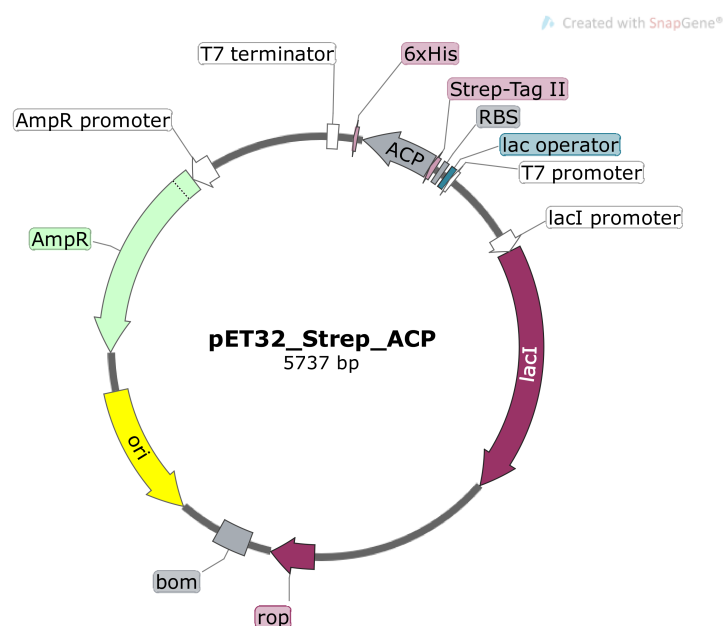


Fig. C.1 Vector Map of pET32 with human ACP (negative control in Kemp assay)

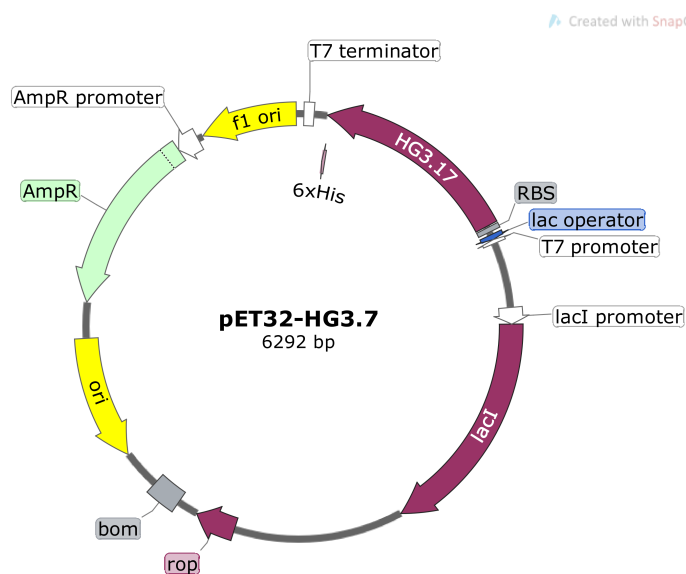


Fig. C.2 Vector Map of pET32 with HG3.17 (positive control in Kemp assay)

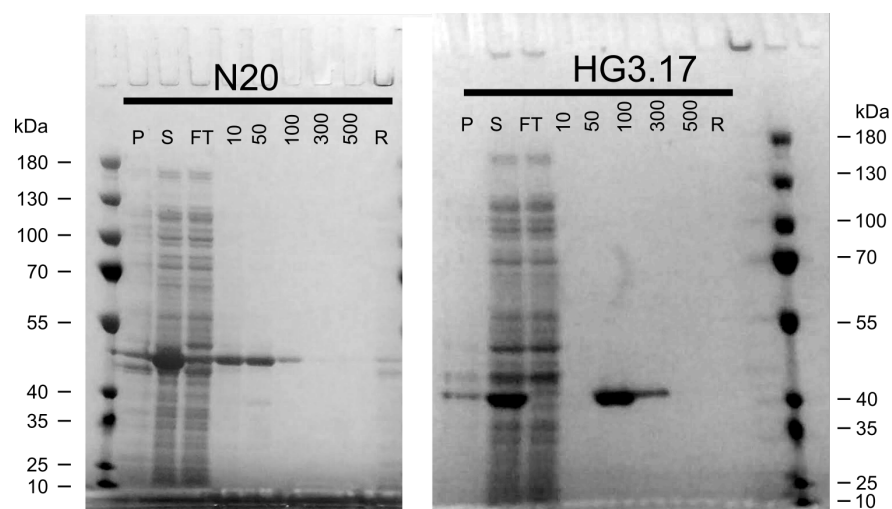


Fig. C.3 Enzymes N20 and HG3.17 were expressed in *E. coli* BL21(Gold) DE3 overnight at 20°C with 400 μ M IPTG. Enzymes were purified by immobilised metal ion affinity chromatography (IMAC). Shown is the elution profile. Fractions eluted at 50 and 100 mM imidazole were combined and concentrated.

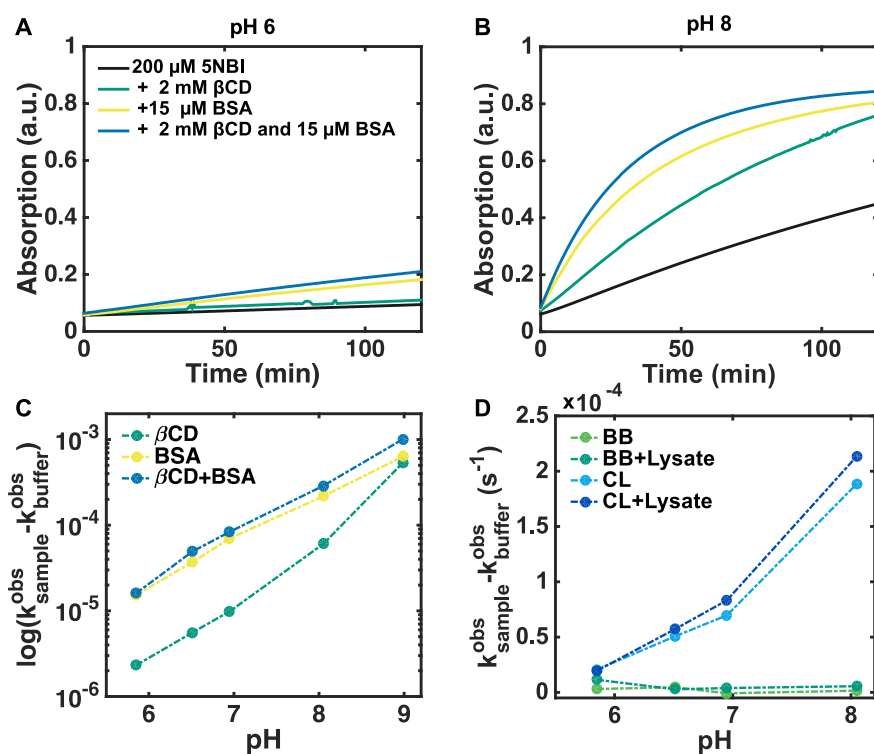


Fig. C.4 A: Absorbance *versus* time for the reaction of 200 μM 5NBI alone or with the indicated reagents in 40 mM sodium phosphate at pH 6 and 100 mM sodium chloride. B: Conditions as in A only at pH 8. C: The observed reaction rate constant for BSA is about 10-fold higher than for βCD up to pH 8. The rate constants were determined by linear fits of the initial velocity of the reaction (up to the first 5 min). D: The effect of BugBuster (BB) and CellLytic (CL) on the reaction. BugBuster did not affect the reaction, where CellLytic led to an acceleration. Addition of cell lysate (average of three measurements) did not further affect the reaction in either case. Error bars were omitted for clarity. All reaction rates k^{obs} were obtained from linear initial velocity fits (slope of first 1 to 5 min of reaction).

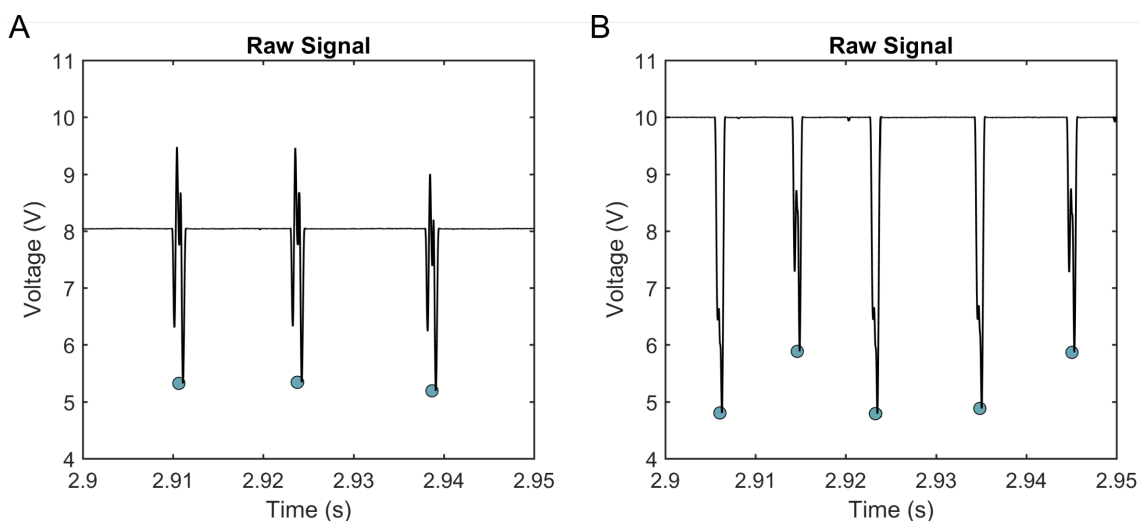


Fig. C.5 Shown is the use of tartrazine to offset the absorbance signal. *A*: If buffer without a dye to offset the absorbance signal is used, a strong edge effect is visible when a droplet enters and leaves the detection area. Furthermore, the signal at the centre of the droplet (corresponding to the transmission of the droplet contents) is very similar to the baseline, which requires the lowering of the incoming light intensity to avoid saturation effects. These effects lower the observed difference between droplets with and without product. *B*: Addition of 3 mM tartrazine increases the absorbance of all droplets, masking the edge effect and lowering the signal below the baseline. Thus, the incoming light intensity can be increased for more sensitive detection.

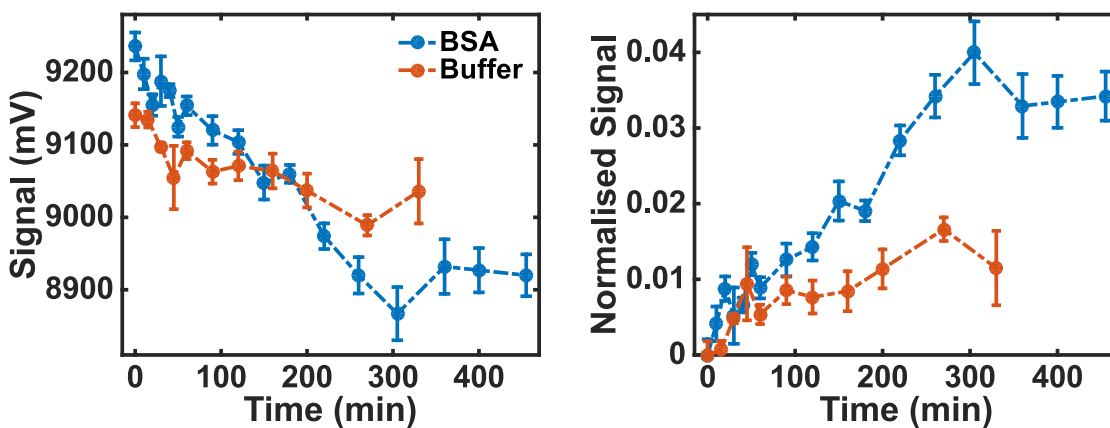


Fig. C.6 The Kemp elimination catalysed by BSA and buffer in microdroplets. *Left*: Average signal *versus* time for droplets containing 50 μM BSA or buffer only with each 500 μM 5NBI (conditions as in Table 4.2, No. 5). The buffer catalysed reaction starts at a lower signal than the BSA catalysed reaction. *Right*: Data normalised according to $S_n(t) = 1 - \frac{S(t)}{S_0}$ revealed BSA catalysis over background. Error bars are the standard deviation of the distribution of the mean droplet values. The dotted lines guide the eye.

HG3.17	ATGGCAGAAG	CAGCACAGAG	CATTGATCAG	CTGATTAAAG	CACGTGGTAA	AGTTTATTTT	GGTGTGGCAA	CCGATCAGAA	TCGTCTGACC	ACCGGTAAAA	ATGCAGCAAT	TATTTAAAGC	GATTT	125
E1	125
E2	125
E3	125
E4	125
E5	125
E6	125
E8	125
E9	125
E10	125
HG3.17	TGGTATGGTT	TGGCCGGAAG	AAAGCATGCA	GTGGGATGCA	ACCGAACC GA	GCCAGGGCAA	TTTTAATTTT	GCGGGTGCA	ATTATCTGGT	TAATTGGGCA	CAGCAGAATG	GTAAACTGAT	TGGTG	250
E1	250
E2	250
E3	250
E4	250
E5	250
E6	250
E8	250
E9	250
E10	250
HG3.17	CAGGTTGTCT	GGTTTGGCAT	AATTTTCTGC	CGAGCTGGGT	GAGCAGCATT	ACCGATAAAA	ATACCTGTAT	TAATGTGATG	AAAAATCATA	TTACCCACCT	GATGACCCGC	TATAAAGGCA	AAATT	375
E1	375
E2	375
E3	375
E4	375
E5	375
E6	375
E8	375
E9	375
E10	375
HG3.17	CGTACTGGG	ATGTTTGGGG	TGAAGCCTTT	AATGAAGATG	GTAGCCTGGG	TCAGATGTGT	TTTCTGAATG	TTATTGGTGA	AGATTATATT	CCGATTGGCT	TTGAGACCGC	CCGTGCCGCC	GATCC	500
E1	500
E2	500
E3	500
E4	500
E5	500
E6	500
E8	500
E9	500
E10	500
HG3.17	GAATGCAAAA	CTGTATATTA	TGGATTATAA	TCTGSAATAGT	GCAAGCTATC	CGAAAAACCA	GGCAATTGTT	AATCGTGTTA	AACAGTGGCG	TGCAGCCGGT	GTTCCGATTG	ATGGTATTGG	TTCAC	625
E1	625
E2	625
E3	625
E4	625
E5	625
E6	625
E8	625
E9	625
E10	625
HG3.17	AGATGCATCT	GAGCGCAGGT	CAGGGTGCAG	GCGTTTCTGCA	GCACTGCGC	CTGCTGGCAA	GTGCAGGTAC	CCCGGAAGTT	AGTATTCTGA	TGCTGGATGT	TGCCGGTGCA	AGTCCGACCG	ATTAT	750
E1	750
E2	750
E3	750
E4	750
E5	750
E6	750
E8	750
E9	750
E10	750
HG3.17	GTTAATGTTG	TTAATGCATG	TCTGAATGTT	CAGAGCTGTG	TTGGTATTAC	CGTTATGGGT	GTTCAGAGAT	CGGATAGCGC	ATTTGCAAGC	AGCACCCCGC	TGCTGTTTGA	TGGTAATTTT	AATCC	875
E1	875
E2	875
E3	875
E4	875
E5	875
E6	875
E8	875
E9	875
E10	875
HG3.17	GAAACCGGCA	TATAATGCAA	TTGTTCAGAA	TCTGCAGCAG	GGTAGCCTCG	AGGGATCCCA	CCATCACCAT	CACCATTA	954					
E1	954					
E2	954					
E3	954					
E4	954					
E5	954					
E6	954					
E8	954					
E9	954					
E10	954					

Fig. C.7 Sequences of nine randomly picked clones from the epPCR library of HG3.17. Amino acid mutations are highlighted in blue, dots indicate no change.

HG3.17	ATGGCAGAAG	CAGCACAGAG	CATTGATCAG	CTGATTAAAG	CACGTGGTAA	AGTTTATTTT	GGTGTGGCAA	CGATCAGAA	TCGTCTGACC	ACCGGTAAAA	ATGCAGCAAT	TATTAAAGCC	GATTT	125
Del1	122
Del2	125
Del3	122
Del4	125
Del5	125
Del6	125
Del7	125
Del8	125
Del9	125
Del10	125
HG3.17	TGGTATGTT	TGGCCGGAAG	AAAGCATGCA	GTGGGATGCA	ACCGAACCAG	GCCAGGGCAA	TTTAAATTTT	GCAGGTGCGA	ATTATCTGGT	TAATTGGGCA	CAGCAGAATG	GTAAACTGAT	TGGTG	250
Del1	247
Del2	250
Del3	247
Del4	250
Del5	250
Del6	250
Del7	250
Del8	250
Del9	250
Del10	250
HG3.17	CAGGTGTGCT	GGTTTGGCAT	AATTTTCTGC	CGAGCTGGGT	GAGCAGCATT	ACCGATAAAA	ATACCTCTGAT	TAATGTGATG	AAAAATGATA	TTACCAACCT	GATGACCCGC	TATAAAGGCA	AAATT	375
Del1	372
Del2	375
Del3	372
Del4	375
Del5	372
Del6	375
Del7	375
Del8	372
Del9	375
Del10	375
HG3.17	CGTACCTGGG	ATGTTGTGGG	TGAAGCCTTT	AATGAAGATG	GTAGCCTGCG	TCAGAATGTT	TTTCTGAATG	TTATTGGTGA	AGATTATATT	CCGATTGCCT	TTCAGACCCG	CCGTGCCGCC	GATTC	500
Del1	497
Del2	500
Del3	497
Del4	500
Del5	497
Del6	497
Del7	500
Del8	497
Del9	500
Del10	497
HG3.17	GAATGC AAAA	CTGTATATTA	TGGATTATAA	TCTGGATAGT	GCAAGCTATC	CGAAAACCCA	GGCAATTGTT	AATCGTGTTA	AACAGTGGCG	TGCAGCCGGT	GTTCCGATTG	ATGGTATTGG	TTCAC	625
Del1	622
Del2	625
Del3	625
Del4	625
Del5	622
Del6	622
Del7	622
Del8	622
Del9	625
Del10	622
HG3.17	AGATGCATCT	GAGCGCAGGT	CAGGGTGCAG	CGCTTCTGCA	GGCACTGCCG	CTGCTGGCAA	GTGCAGGTAC	CCCCGAAAGT	AGTATTCTGA	TGCTGGATGT	TGCCGGTGCA	AGTCCGACCG	ATTAT	750
Del1	747
Del2	747
Del3	747
Del4	747
Del5	747
Del6	747
Del7	747
Del8	747
Del9	750
Del10	747
HG3.17	GTTAATGTTG	TTAATGCATG	TCTGAATGTT	CAGAGCTGTG	TTGGTATTAC	CGTTATGGGT	GTTGCAGATC	CGGATAGCGC	ATTTGCAAGC	AGCACCCCGC	TGCTGTTTGA	TGGTAATTTT	AATCC	875
Del1	872
Del2	872
Del3	872
Del4	872
Del5	872
Del6	872
Del7	872
Del8	872
Del9	875
Del10	872
HG3.17	GAAACCGSCA	TATAATGCAA	TTGTTTCAGAA	TCTGCAGCAG	GGTAGCctcg	agcaccacca	ccaccaccac	tgagatccgg	ctgcttaa					952
Del1	959
Del2	959
Del3	959
Del4	959
Del5	959
Del6	959
Del7	959
Del8	959
Del9	959
Del10	959

Fig. C.8 Sequences of ten randomly picked clones from the deletion library of HG3.17. Amino acid deletions are highlighted in blue, dots indicate no change.

HG317	ATGGCAGAAG	CAGCACAGAG	CATTGATCAG	C---TGATTA	AAGCACGTGG	TAAAGTTTAT	TTTGGTGTGG	CAACCGATCA	GAATCGTCTG	ACCACCGGTA	AAAATGCAGC	AATTATTAAA	GCCGA	122
Ins1														122
Ins2														122
Ins3														122
Ins4														122
Ins5														122
Ins6														122
Ins7														122
Ins8														125
Ins9														122
Ins10														122
HG317	TTTTGGTATG	GTTTGGCCGG	AAGAAAGCAT	GCAGTGGGAT	GCAACCGAAC	CGAGCCAGGG	CAATTTTAAAT	TTTGGCGGTG	CAGATTATCT	GGTTAATTGG	GCACAGCAGA	ATGGTAAACT	GATTG	247
Ins1														247
Ins2														247
Ins3														247
Ins4														247
Ins5														247
Ins6														247
Ins7														247
Ins8														250
Ins9														122
Ins10														247
HG317	GTGCAGGTTG	TCTGGTTTGG	CATAATTTTC	TGCCGAGCTG	GGTGAGCAGC	ATTACCGATA	AAAATACCGT	GATTAATG-T	GATGAAAAAT	CATATTACCA	CCCTGATGAC	CCGCTATAAA	GGCAA	371
Ins1														372
Ins2														371
Ins3														371
Ins4														371
Ins5														371
Ins6														371
Ins7														371
Ins8														374
Ins9														371
Ins10														371
HG317	AATTCGTACC	TGGGATGTTG	TGGGTGAAGC	CTTTAATGAA	GATGGTAGCC	TGCGTCAGAA	TGTTTTCTGT	AATGTTATTG	GTGAAGATTA	TATTCGGATT	GCCTTTCAGA	CCGCCCGTGC	CGCCG	496
Ins1														497
Ins2														496
Ins3														496
Ins4														496
Ins5														496
Ins6														496
Ins7														496
Ins8														499
Ins9														496
Ins10														496
HG317	ATCCGAATGC	AAAACGTGAT	ATTATGGATT	ATAATCTGGA	TAGTGCAAGC	TATCCGAAAA	CCCAGGCAAT	TGTTAA---T	CGTGTTAAAC	AGTGGCGTGC	AGCCGGTGTT	CCGATTGATG	GTATT	618
Ins1														619
Ins2														618
Ins3														618
Ins4														618
Ins5														618
Ins6														621
Ins7														621
Ins8														618
Ins9														618
Ins10														618
HG317	GGTTCACAGA	TGCATCTGAG	CGCAGGTCAG	GGTGC---AG	GCCTTCTGCA	GGCACTGCCG	CTGCT---GG	CAAGTGC--A	GGTACCCCGG	AAGTTAGTAT	TCTGATGCTG	GATGTTGCCG	GTGCA	735
Ins1														736
Ins2														735
Ins3														738
Ins4														735
Ins5														735
Ins6														738
Ins7														735
Ins8														738
Ins9														737
Ins10														737
HG317	AGTCCGACCG	ATTATGTAA	TGTTGTAAAT	---GCATGTC	TGAATGTTCA	GAGCTGTGTT	GGTATTACCG	TTATGGGTGT	TGCAGATCCG	GATAGCGCAT	TTGCAAGCAG	CACCCCGCTG	CTGTT	857
Ins1														858
Ins2														857
Ins3														860
Ins4														860
Ins5														857
Ins6														860
Ins7														857
Ins8														860
Ins9														860
Ins10														859
HG317	TGATGGTAAT	T---TAAATCC	GAAACCGG--	---CATATAATG	CAATTGTTCA	GAATCTGCAG	CAGGGTAGCc	tcgagcacca	ccaccaccac	cacttga				948
Ins1														949
Ins2														950
Ins3														951
Ins4														951
Ins5														951
Ins6														951
Ins7														948
Ins8														951
Ins9														951
Ins10														950

Fig. C.9 Sequences of ten randomly picked clones from the insertion library of HG3.17. Amino acid insertions are highlighted in blue, dots indicate no change.

¹³C-NMR (400 MHz, DMSO-d₆)

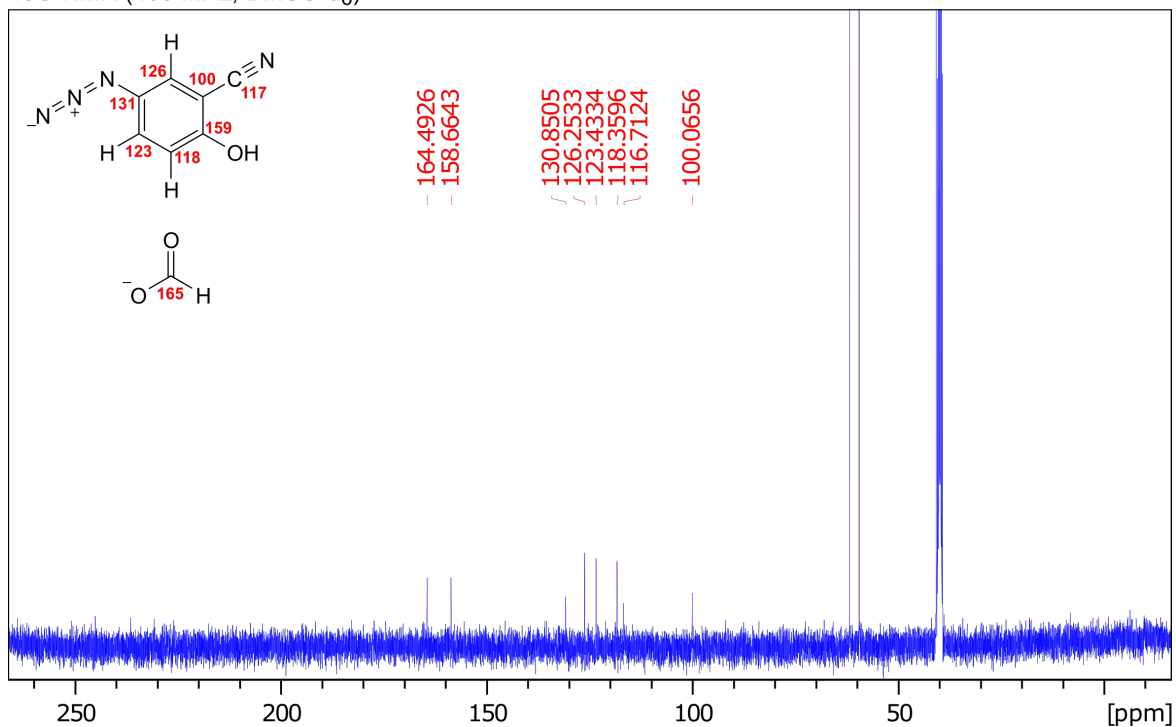


Fig. D.1 ¹³C-NMR of **6c**.

Appendix D

Supplementary Data for Chapter 5

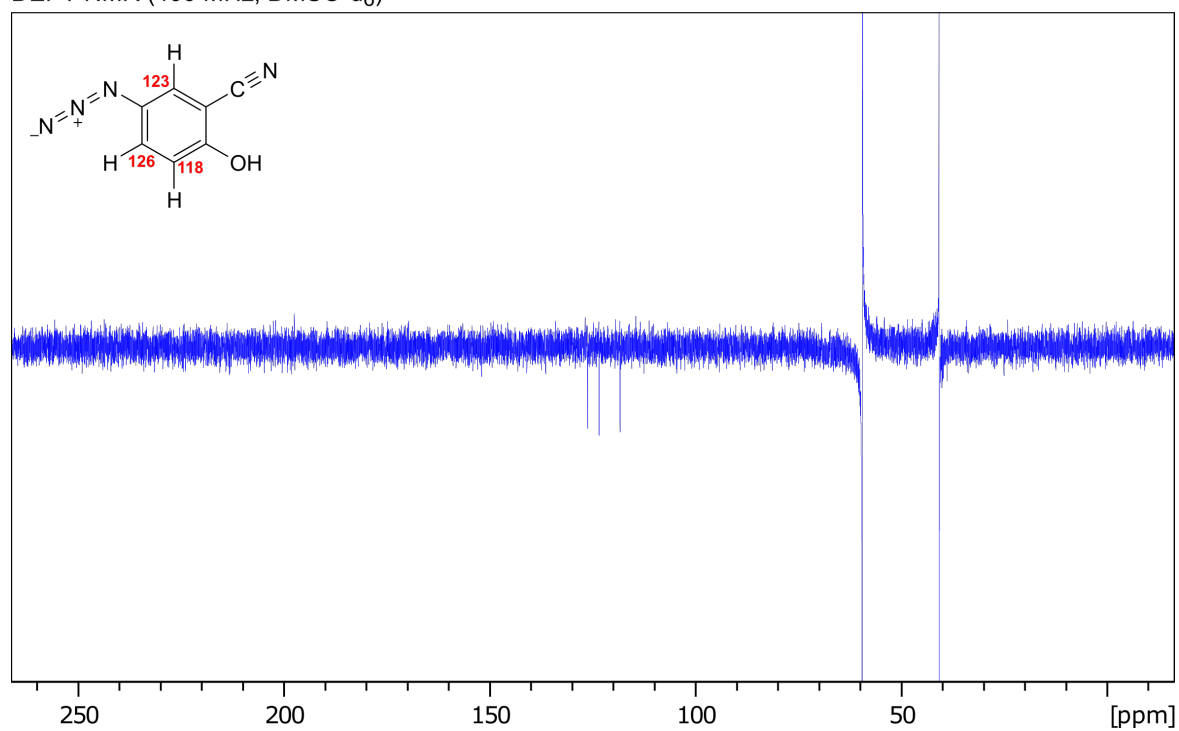
DEPT-NMR (400 MHz, DMSO-d₆)

Fig. D.2 DEPT-NMR of **6c** (negative signals are from carbon atoms connected to a hydrogen atom via one σ -bond).

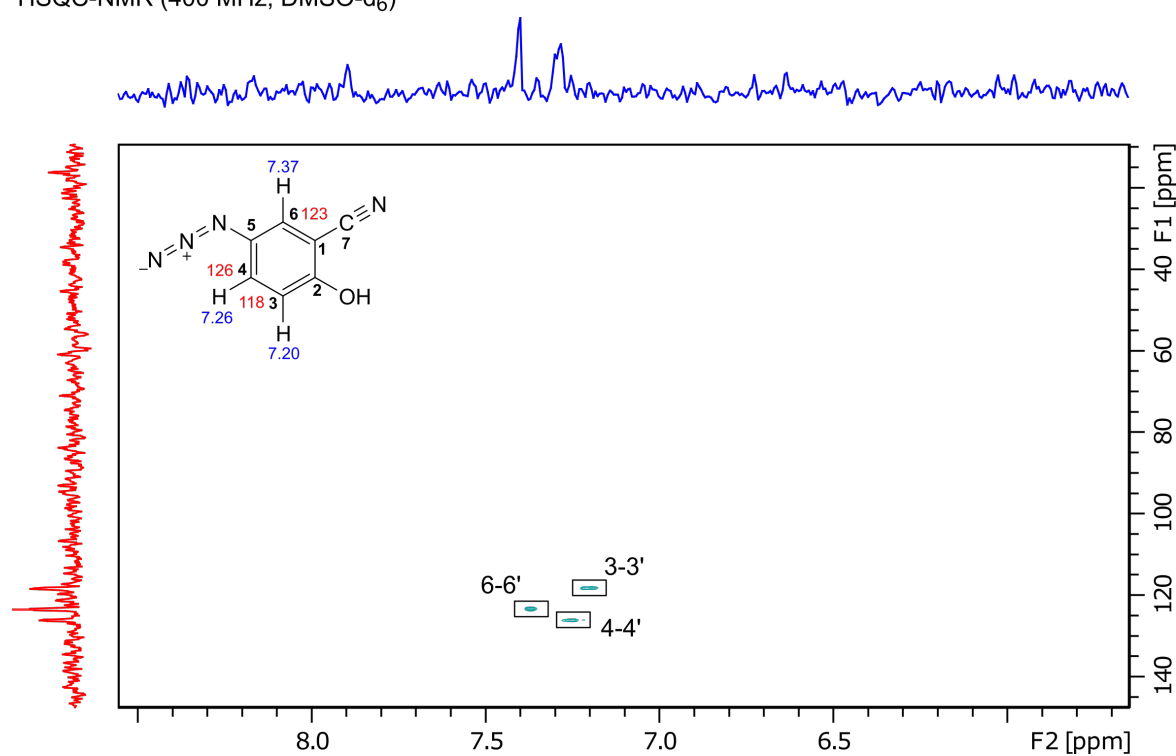
HSQC-NMR (400 MHz, DMSO-d₆)

Fig. D.3 HSQC-NMR of **6c** (HSQC shows the coupling between carbon and hydrogen atoms connected via one σ -bond).

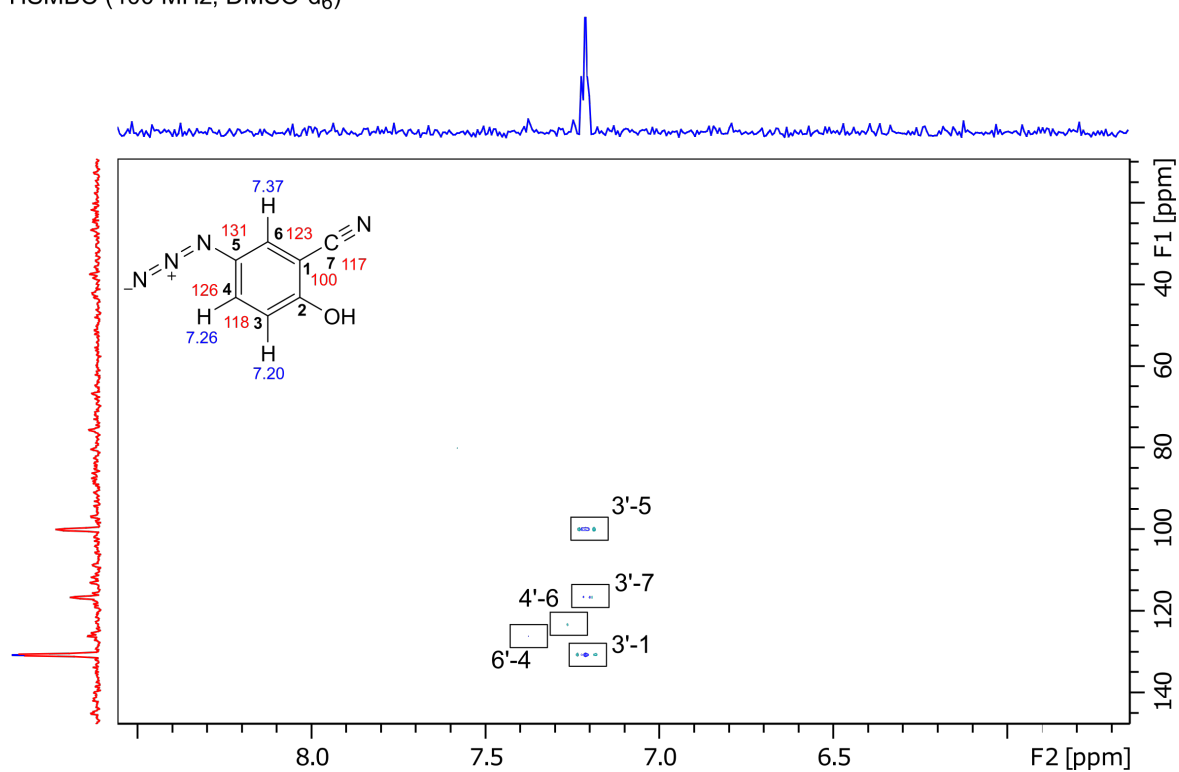
HSMBC (400 MHz, DMSO-d₆)

Fig. D.4 HMBC-NMR of **6c** (HMBC shows the coupling between carbon and hydrogen atoms two or three σ -bonds away from each other).

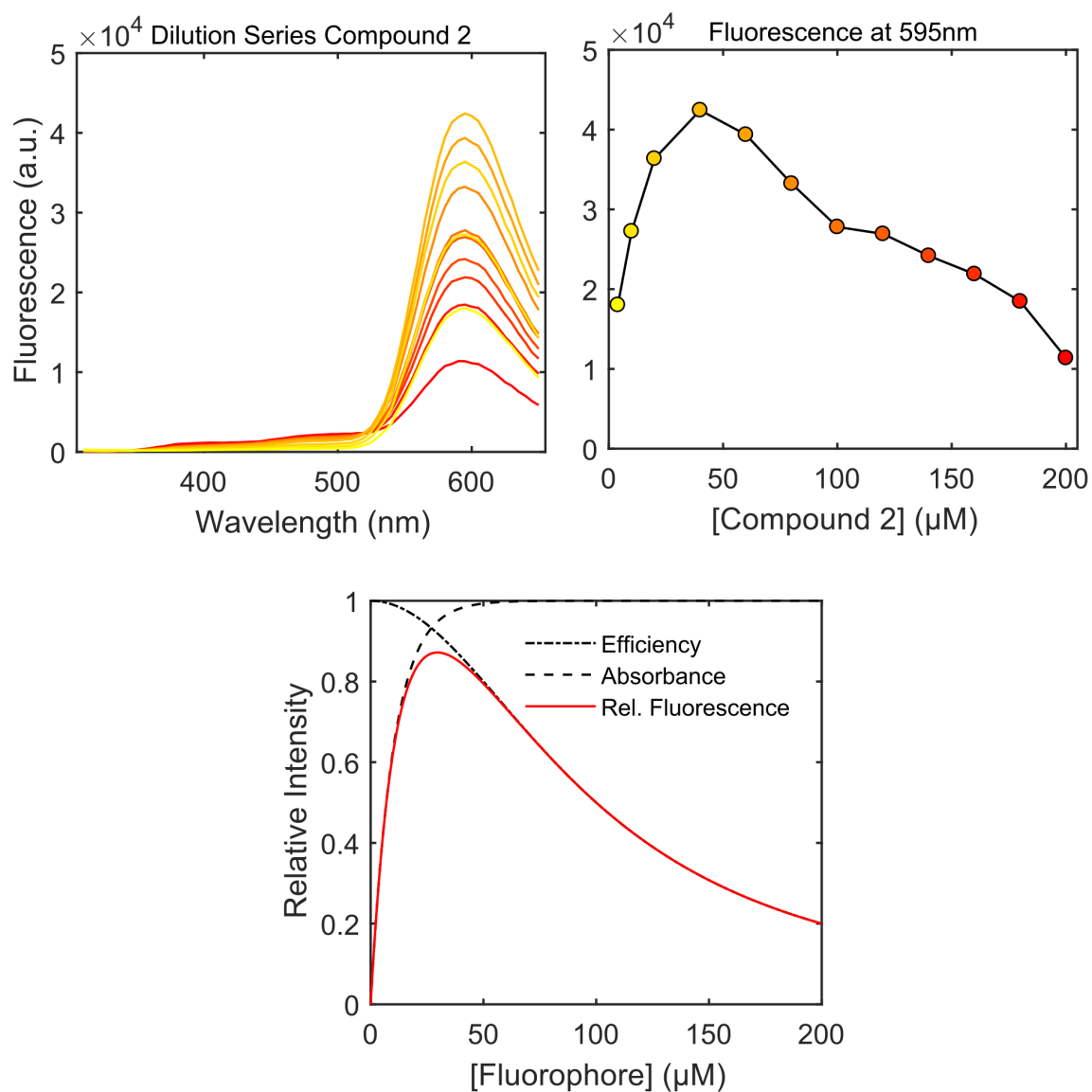


Fig. D.5 The fluorescence of **6c** increased up to 40 μM above which the fluorescence decreased again. This self-quenching was indicative of resonance energy transfer between individual molecules, a behaviour observed for many fluorophores including fluorescein. The bottom panel shows the two competing effect on relative fluorescence intensity: increasing absorbance and decreasing efficiency according to [267].

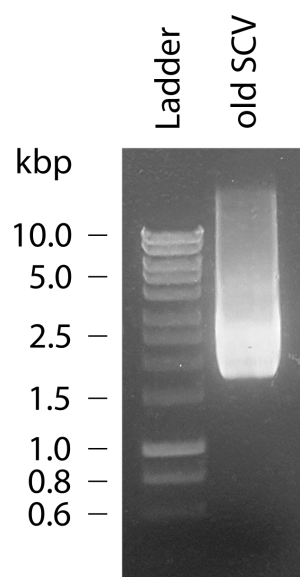


Fig. D.6 Gel-electrophoresis of the old SCV library (undigested plasmids) shows a smear indicating degradation of the library.

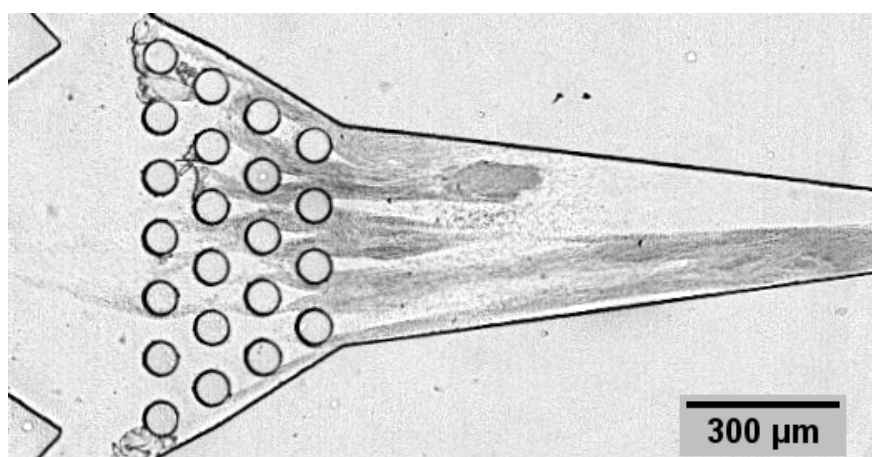


Fig. D.7 If *E. coli* cells are grown at high density for prolonged time on agar plates, they secrete polysaccharides to protect themselves. Even after thorough washing, individual cells are sticky enough to start forming a biofilm (dark grey strings) during droplet generation. Strings of cells can break off, which causes the co-encapsulation of multiple cells.